Bias propagation in Bayesian meta-analysis: A simulation study

Finley Draffin-Jacquin

Minerva University

Prof. Jon Wilkins

Mar 30, 2024

14206 Words

Bias propagation in Bayesian meta-analysis: A simulation study

Executive Summary

The study explores the resilience of Bayesian meta-analysis against various forms of bias, specifically prior selection, and compares the impact of this bias to the widely recognised phenomenon of p-hacking in frequentist statistics. I aimed to empirically evaluate the robustness of Bayesian hypothesis testing under different prior distributions and true effect sizes by using simulated datasets and assessing the extent to which prior selection influenced the outcomes of statistical analyses both frequentist and Bayesian.

The core of the investigation involved the computational simulation of 50 studies (Figure 1), each with a placebo and control group, with varying differences in means $\Delta\mu$ to represent different effect sizes. Bayes' factors were then calculated under five distinct prior distributions for each $\Delta\mu$ group.

The findings indicate a nuanced relationship between prior selection and the resulting Bayes' factors. Contrary to initial expectations, the selection of biased (narrower) priors did not significantly skew the results towards either hypothesis, suggesting an inherent resilience of the Bayesian approach to the manipulations akin to p-hacking. This was particularly evident for larger effect sizes, where Bayesian methods consistently favored the alternative hypothesis, aligning with the true simulated effects.

However, for smaller effect sizes, the influence of prior selection became more pronounced, albeit not to the extent anticipated. The study further juxtaposed these Bayesian outcomes against traditional frequentist tests, revealing a comparable inability of both methodologies to detect very small effect sizes, underscoring a common limitation across statistical paradigms.

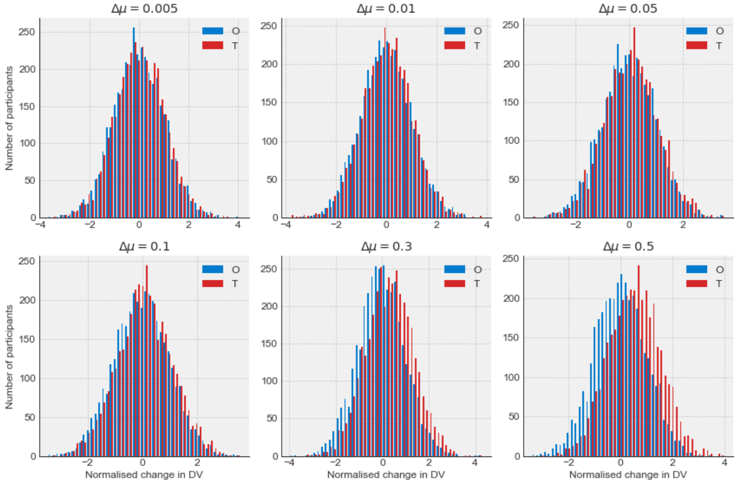

Figure 1

Aggregated data from the simulated studies

The implications of these findings are multifaceted. Firstly, they bolster the argument for the potential of Bayesian methods as a solid defense against statistical manipulation. Secondly, they highlight the importance of considering effect sizes and sample sizes in the design and interpretation of both Bayesian and frequentist analyses. Lastly, the study underscores the need for further empirical investigations into the effects of other elements of Bayesian analysis (such as likelihood selection) on bias propagation.

Introduction

In the courtroom of science, hypothesis tests are the jury. Unfortunately for the plaintiff, they are a deeply flawed and biased panel. Despite being used to assess evidence in favour of an alternative hypothesis, null-hypothesis tests offer no direct evidence to the plausibility of an alternative hypothesis (Berkson, 1943; Hubbard et al., 2003; Ioannidis, 2005; Schervish, 1996). Moreover, they are highly sensitive to bias, and despite researchers imposing Type I error thresholds α, in the face of publication bias and p-hacking, those error rates can be far higher than specified αs (Bartoš & Schimmack, 2022; Fanelli, 2011; Franco et al., 2014; McShane & Gelman, 2022; Rosenthal, 1979; Rothstein et al., 2005; Simonsohn et al., 2014b; Thornton & Lee, 2000). Bayesian approaches not only directly probe an alternative hypothesis (Wagenmakers, 2007; Wagenmakers & Grünwald, 2006; Wei et al., 2022), but are also proclaimed to be less prone to bias via researcher freedom (Ioannidis, 2005, 2008). However, the increased subjectivity of Bayesian methods remains a criticism (Dienes, 2016; Ioannidis, 2005; Wagenmakers & Grünwald, 2006). I perform a simulation study that calculates Bayes' factors for a difference of means test under priors of varying certainty and degrees of bias to further investigate this claim.[1]

Literature Review

Hypothesis-testing: an overview

A typical hypothesis test begins with a null hypothesis H_0, which typically states that in whatever is being tested, there is no relationship. That the difference between two test groups is 0, that the slope of a regression line is 0 and so forth. Formally, H_0 states that for two groups A and

[1] #ORGANISATION: The primary goal of this paper is to present Bayesian methods as a stronger, more robust statistical approach to statistical testing than frequentist approaches. I open the paragraph, and the paper entire, with a metaphor. Comparing a courtroom jury to a hypothesis test immediately shows the reader why hypothesis testing is important, and why we ought to care about them being flawed. This sets a critical tone that I move on to justify when I then present the theoretical and practical flaws of frequentist approaches. This provides the reader with the necessary information to understand why frequentist approaches are weak. By then pivoting to Bayesian methods, and by describing how they directly address the flaws just described, I present the claim that Bayes offers a better statistical approach to hypothesis testing. By presenting the conflicting evidence to this end, I motivate my research question. Thus the organisation of this paragraph provides a microcosm of the entire literature review, succinctly explaining and justifying the purpose of the study.

B defined by a parameter θ, $\theta_A - \theta_B = 0$. As they are employed today, if the evidence in favour of H_0 is weak, it is turned town.

A researcher begins with a null hypothesis H_0. Frequently, this is the hypothesis of there being no effect. For instance, for a hypothesis about a parameter θ, H_0 is often that $\theta = 0$. θ may be any parameter: a mean, a difference in means, a regression coefficient etc. There are two formal perspectives on frequentist hypothesis testing: the Fisherian and the Neyman-Pearson. Both compute a p value, but each interprets it differently. Fisher's p is $P(\theta|H_0)$: the smaller the p the smaller the likelihood that the data were generated under the null hypothesis and the weaker the evidence in favour of H_0 (Fisher, 1971). In Fisher's approach, H_0 is rejected in favour of H_0^c, an infinite set of alternative hypotheses (Hubbard et al., 2003). If $p < \alpha$, with α frequently and arbitrarily set to 0.05, Fisher suggests that we reject H_0 (Fisher, 1971).

In the Neyman-Pearson approach, H_0 is pitted against some specifically defined alternative hypothesis H_A. If H_0 is a point hypothesis ($\theta = 0$), H_A may be one or two sided (either $\theta \neq 0$ or $|\theta| > 0$. Since H_A is now explicitly specified, researchers can make Type I and II errors, either falsely rejecting H_0 (I) or falsely accepting it (II). Researchers must specify the maximum tolerable Type I and II error rates α and β (Neyman & Pearson, 1928). This is in contrast to the Fisherian approach, where a researcher specifies the minimum tolerable evidence for H_0 (Fisher, 1971).

Unfortunately, most research sits at an awkward marriage of the Fisher and Neyman-Pearson approaches that can easily lead authors to misinterpret their results (Berkson, 1943; Hubbard et al., 2003; Schervish, 1996). Comparing p to α tempts one to see p as an observed Type I error rate, which it fundamentally is not (Schervish, 1996); p is still only a measure of evidence for H_0. As Fisher stressed, p has nothing to do with alternative hypotheses.[2]

[2] #PROBABILITY: In this section, I explore the complexities and nuances of probability as it pertains to hypothesis testing in scientific research. Specifically, I address the conceptual differences between Fisherian and Neyman-Pearson approaches to calculating and interpreting p-values. I aim to elucidate how probabilities can be misinterpreted, and how such misinterpretations can have a detrimental impact on the validity and reliability of research outcomes. The intention is to shed light on the ambiguity and challenges associated with the use of probability in hypothesis testing, an element that is often assumed to be straightforward but is riddled with conceptual intricacies.

Bayesian Hypothesis Testing

Bayesian methods present an alternative to null hypothesis testing that obviates the theoretical impracticalities of null hypothesis testing (Wagenmakers & Grünwald, 2006; Wei et al., 2022), namely that they do not explicitly probe H_A.

In the Bayesian approach one computes the Bayes' factor (B), the likelihood ratio of H_A's probability to H_0's given the data. B provides direct evidence as to the plausibility of one hypothesis over another. In the case that the alternative is even less likely than an already unlikely H_0, the null-hypothesis test would lead the researcher to reject H_0, while the Bayesian test would concretely reveal H_0 as the more likely of the two; the Bayesian approach evaluates multiple hypotheses (Wagenmakers, 2007).

Mathematical description

The fundamental object in Bayesian hypothesis testing is the Bayes' factor B:

$$B = \frac{P(D|H_0)}{P(D|H_1)} \tag{1}$$

B is the ratio of the evidence for one hypothesis over another as given by the data. Under a hypothesis H_α, $P(D|H_\alpha)$ is the marginal likelihood over all parameters:

$$P(D|H_\alpha) = \int P(D|\theta, H_\alpha)P(\theta)d\theta \tag{2}$$

Where $P(D|\theta, H_\alpha)$ is the joint likelihood of the data under H_α and the model parameters θ and $P(\theta)$ the prior belief on the parameters. Note that this is the same formula used for model selection. In Bayesian hypothesis testing, hypotheses are modelled with specific θ values and B quantifies which hypothesis (or rather, model of a hypothesis) is more likely.[3]

For models with many parameters, this is a high-dimensional integral, often with no

[3] #PROBABILITY: I describe how the Bayes' factor is computed and explain how the Bayesian approach to hypothesis testing differs from frequentist null hypothesis testing. I explain how the Bayesian approach improves on the flaws of null hypothesis testing in terms of comparing hypotheses and assessing their relative plausibility.

analytical solution. For such problems, Markov Chain Monte Carlo (MCMC) sampling is a common approach.

Bias in hypothesis testing

Though hypothesis testing is meant to provide, at a minimum, an objective assessment of the data's plausibility under the null hypothesis, this objectivity quickly crumbles at the hands of human bias and error (Bartoš & Schimmack, 2022; Fanelli, 2011; Franco et al., 2014; Simonsohn et al., 2014b; Thornton & Lee, 2000). Publication bias leads to increased Type I error rates and overestimated effect sizes (McShane & Gelman, 2022; Rosenthal, 1979; Rothstein et al., 2005). *p*-hacking, whether conscious or unconscious, further inflates the Type I error rate by manipulating *p* values into a 'publishable' range while, finally, the belief that *p* values and Type I error rates are the same leads authors to underestimate the impacts of their biases.

Publication bias

Publication bias, a pervasive issue across scientific disciplines, significantly distorts the scientific literature, leading to an increase in Type I error rates and an upward bias in effect sizes (McShane & Gelman, 2022; Rosenthal, 1979). This bias predominantly arises from the preferential publication of studies with positive results, a trend that has been intensifying over recent decades (Fanelli, 2011; Pautasso, 2010). Fanelli observed a notable increase in the publication of positive findings from 1990 to 2007, particularly in the social and applied sciences, which coincides with the periods of significant replication crises in these fields (Fanelli, 2011; Joober et al., 2012). This situation is often exacerbated by the 'file drawer problem,' where non-significant studies remain unpublished, skewing meta-analyses towards falsely inflated effect sizes (Rosenthal, 1979). While various factors contribute to publication bias, including editorial preferences and researchers' reluctance to submit negative results, the root of the problem often lies in the complex interplay between researchers, journals, and funding agencies, each influenced by distinct motivations and biases (Johnson & Dickersin, 2007; Rothstein et al., 2005; Thornton

& Lee, 2000).[4]

Editors. Many point the finger at editors (Fanelli, 2011; Thornton & Lee, 2000), accusing them of favouring positive results. Indeed, considering the status and newsworthiness of positive results over negative ones (Koren & Klein, 1991), and the power of confirmation bias, it is not unreasonable to assume that editors favour positive results (Johnson & Dickersin, 2007). Yet, there is little statistical evidence to suggest that editorial bias exists (Olson et al., 2002; Rothstein et al., 2005; Seward, 1999).

Funding agencies. On a higher level, funding agencies present another source of bias (Johnson & Dickersin, 2007). A notable proportion of scientific research is either privately funded or has profit potential (Aghion et al., 2008; Eisenberg, 1988; Olson et al., 2002). Funding from private sponsors often comes with an agenda, whether that be financial, political or otherwise, and this agenda can often direct what findings are published (Eisenberg, 1988; Johnson & Dickersin, 2007). This is especially problematic in the pharmaceutical industry, where positive results are the first sign of a highly profitable product (Lexchin et al., 2003). Numerous studies have demonstrated that industry funding is a significant predictor of a positive outcome within the industry (Davidson, 1986; Easterbrook et al., 1991; Finucane & Boult, 2004; Yaphe et al., 2001), with the practice sometimes leading to the prescription of ineffective drugs (see for example Roth et al. (2004). In a survey of 372 articles investigating the efficacy of drugs treating schizophrenia, none of the 124 industry-sponsored studies reported negative results, with the rate of positive results markedly higher within the industry-funded selection (Procyshyn et al., 2004). There is a notable caveat here, which is that industry studies may be of higher quality (Krimsky, 2013) and that only candidate drugs internally determined to have potential are subject to academic study (Yaphe et al., 2001), though the evidence is conflicting (Easterbrook et al., 1991).

[4] #ORGANISATION: This paragraph aims to present a concise overview of the section to follow. I open by stating the effect of publication bias on error rates and effect sizes and note that this is a problem across disciplines, establishing therefore its weight as an issue. I then briefly describe the causes of publication bias across levels (especially effective since the whole section takes a #LEVELSOFANALYSIS perspective. My concluding sentence emphasises the importance of this perspective: publication bias is difficult to prevent due to it arising not only within levels but also between them.

The problem extends as well outside of pharmaceuticals (Krimsky, 2013), with funding sources predicting positive findings in topics from climate change (Almond et al., 2022; Farrell, 2016; Michaels, 2008), to tobacco (Barnes & Bero, 1996; Hendlin et al., 2019; Malone, 2013) and chemical toxicity (Bolt, 2011), to name a few. Funding agencies with a vested interest in obtaining certain results (tobacco, pharmaceuticals, the chemical industry etc.) will gravitate to funding those researchers and studies with a greater potential for positive results. Researchers, who have an interest in being funded, may also succumb to manipulating their study to appease the funding source, creating a vicious cycle of bias that seeks positive results.[5]

Researchers. It is not just that negative results are *rejected* by journals, but rather also that they are never submitted (Johnson & Dickersin, 2007; Olson et al., 2002; Seward, 1999). Between data generation and publication lies a rift wherein a researcher decides whether to write up and submit a study, or move on to a new idea (Rothstein et al., 2005). Dickersin notes that a major reason researchers decide not to continue a study at this stage is that the results were negative or uninteresting (1997). A belief also pervades amongst researchers that negative results are more likely to be rejected by journals (Johnson & Dickersin, 2007; Joober et al., 2012; Rothstein et al., 2005), yet this tends not to be the case (Olson et al., 2002; Rothstein et al., 2005; Seward, 1999).

The problem is exacerbated by the fact that researchers' success is driven largely by publications, citations, and public attention (Joober et al., 2012). Positive results are more likely to be newsworthy (Koren & Klein, 1991). Furthermore, the majority of highly cited studies are published in an exclusive subset of journals (Callaham et al., 2002; Ioannidis, 2006), and positive results may be more likely to be cited (Gøtzsche, 1987; Kjaergard & Gluud, 2002; Rothstein et al., 2005) (see Callaham et al. (2002) and Christensen-Szalanski and Beach (1984) for evidence to the contrary), especially within these prestigious journals (Ioannidis, 2006). In addition to

[5] #NS125-WORKINCONTEXT: I provide multiple real-world examples of publication bias and its negative effects. This motivates why we ought to care about publication bias in the first place: the fruits of scientific studies affect the products and drugs that people consume. Accurately assessing the extent of this bias helps to prevent false claims from permeating into the economy.

encouraging researchers, citations are also important in identifying studies for meta-analysis (Rothstein et al., 2005). Driven by a desire to get results that are seen, it is unsurprising that researchers shelve their negative results, whether they are right to do so or not.

To understand the systemic nature of publication bias, we must examine the interactions between researchers, journals, and funding agencies (Johnson and Dickersin (2007) and Olson et al. (2002)). Funding agencies with a vested interest in positive results can indirectly influence editorial decisions by pressuring researchers to produce favourable data (Eisenberg (1988) and Lexchin et al. (2003)). Researchers, already operating under the perception that negative results are less publishable, are further incentivised to selectively report outcomes (Dickersin (1997) and Rothstein et al. (2005)). This selective reporting feeds into the academic journals, where, contrary to popular belief, editors may not be the gatekeepers of bias but rather unwitting participants in a cycle that begins with the funding sources (Olson et al. (2002)). The interplay between these levels reveals a more nuanced picture of the publication bias landscape, necessitating a multi-disciplinary and multi-level approach for meaningful reform (Johnson and Dickersin (2007) and Krimsky (2013)).[6]

p-hacking

Researchers' biases significantly impact statistical decisions in studies, particularly through '*p*-hacking,' where data or statistical analyses are manipulated to produce artificially significant results from initially non-significant findings (Head et al., 2015; Simmons et al., 2011; Simonsohn et al., 2014b). Common tactics include adjusting sample sizes and halting analysis upon achieving significance or excluding outliers. Such practices inflate false positive rates, compromising the reliability of individual studies and broader research fields ((Simmons et al., 2011)). Simmons et al. highlighted this issue in their computational study, demonstrating how the

[6] #LEVELSOFANALYSIS: This HC serves to deconstruct and scrutinize the complex issue of publication bias through multiple tiers: funding agencies, researchers, and academic journals. Each level was not only distinctly identified but also rigorously analyzed using evidence from various studies. The necessity for a multi-level approach was justified by showing that a single-level analysis would provide an incomplete understanding of the issue. Further, I went beyond merely identifying these levels by delving into the interactions and interdependencies among them, thereby offering an integrated view.

combination of various degrees of researcher freedom (like selecting dependent variables and reporting results selectively) can escalate the false positive probability to 61%, well above the typical $\alpha = 0.05$ threshold (2011). This effect is exacerbated when coupled with publication bias, particularly in fields with small effect sizes, such as clinical trials, where the distortion of results becomes even more pronounced ((Begg & Berlin, 1988; Friese & Frankenbach, 2020)).

Researchers are often driven to shelve negative results (see Publication Bias). In the face of a fierce belief in a hypothesis, a researcher may instead *p* hack to obtain the conclusion he desires. Often seeking positive results (Olson et al., 2002), pulled too by sunk-costs and strong confirmation bias, and faced with ambiguous statistical decision making (Simmons et al., 2011; Wicherts et al., 2016), *p*-hacking is likely a ubiquitous behaviour, conscious and unconscious.[7]

Meta-analysis

Evaluating a scientific hypothesis or the effect size from a specific intervention requires more than single studies, which are often compromised by publication bias, *p*-hacking, and irreproducibility. Moreover, single studies investigating small effect sizes may lack sufficient power and be especially prone to false negatives (Freiman et al., 1978). Single studies are insufficient evidence for a hypothesis.

The kernel of the meta-analysis is the effect size. Kelley and Preacher (2012, p.140) provide a meaningful definition:

> Effect size: A quantitative reflection of the magnitude of some phenomenon that is used to address a question of interest.

This is a broad definition and there is no single formula for effect size; rather *effect size* represents a category of statistics: Cohen's d, RMSEA or standardised mean difference are all examples of effect sizes (Kelley & Preacher, 2012). A meta-analysis pools together studies that measure the same effect, thereby providing us with a more accurate estimate of the true effect size.

[7] #BIASIDENTIFICATION: I've identified key cognitive biases like confirmation bias and the sunk-cost fallacy as potential drivers behind the practice of *p*-hacking in research. I've also explained how these biases lead researchers to make specific decisions that contribute to the skewing of scientific results.

Meta-analysis aims to correct for these issues (Egger et al., 2009). Typically seeking a single treatment effect estimate, a meta-analysis involves a statistical analysis of findings from independent studies (Egger et al., 2009); it is a study of studies (Lipsey & Wilson, 2001, p.1). A meta-analysis requires that the body of research surveyed be quantitative and investigate similar or comparable relationships and mechanisms, thereby necessitating a sort of "coding" wherein findings from separate studies be selected and standardised for statistical comparison (Lipsey & Wilson, 2001, p.4). By pooling multiple studies together, meta-analyses offer vastly larger effective sample sizes than any of its component studies, and, in principle, the aggregation of studies ought to neutralise the biases present. Unfortunately, as both publication bias and *p*-hacking skew toward positive results, this correction occurs imperfectly.[8]

Meta-analyses are not the only way to survey a body of literature, however. They form a subset of the broader "systematic review," which aims to assess an entire body of literature systematically and transparently (Egger et al., 2009). Narrative reviews have the author qualitatively summarise a body of research and are especially prone to selection bias (Gøtzsche, 1987; Kjaergard & Gluud, 2002; Schmidt & Gøtzsche, 2005). Narrative reviews may even contradict meta-analytical conclusions (Schmidt & Gøtzsche, 2005). Meta-analyses, by providing a more objective assessment of a body of literature, lend more reliable conclusions that may be more rigorously assessed and confirmed.

Publication bias in meta-analyses

Meta-analyses, though intended to mitigate publication bias and *p*-hacking, are not entirely bias-immune (Dickersin, 2005). Publication bias affects study selection (Berlin & Ghersi, 2005) and skews effect estimates (Simonsohn et al., 2015; Sutton, 2005), while biased reporting within studies (Sutton & Pigott, 2005) can also affect meta-analysis outcomes. To address these challenges, researchers have explored various methods, including pre-registration of clinical trials (Berlin & Ghersi, 2005; Simes, 1986) and the adoption of prospective meta-analyses (Berlin &

[8] #BIASMITIGATION: I explain how meta-analyses aim to mitigate publication bias and *p*-hacking by averaging over these biases and offering larger effect sizes. I explain why this is insufficient.

Ghersi, 2005; Clarke & Stewart, 2009; Stewart et al., 2005). However, while these approaches offer potential solutions, they also come with their own set of limitations and concerns.

An unbiased meta-analysis assumes an unbiased selection of studies. When this isn't the case, however, the analysis can produce a biased conclusion (Berlin & Ghersi, 2005). Traditionally, the studies included in a meta-analysis are determined via researcher-defined criteria (Egger & Davey Smith, 2009, pp. 22-27). Selection bias and researcher subjectivity in the study inclusion process can skew a meta-analysis towards a certain finding by excluding studies to the contrary. By pre-registering clinical trials, meta-analysts may become aware of unpublished studies, and produce a less-biased analysis (Berlin & Ghersi, 2005; Simes, 1986). However, pre-registration provides no guarantee that the register itself will not be biased: non-disclosure of negative results remains a problem, especially in industry-sponsored trials (Manzoli et al., 2014). Though regulatory encouragement (e.g. by the FDA) for pre-registration has reduced this issue, it has not eliminated it (Zou et al., 2018). Moreover, *p*-hacking and selective reporting remain issues even in pre-registered trials (Rising et al., 2008; Scott et al., 2015). Again, this undermines the validity of the research and leaves even a meta-analysis of pre-registered trials prone to bias and erroneous conclusions. In a medical context, this can create misleading clinical guidelines (Egger et al., 2009, pp. 9-10), adoption of ineffective treatments (Roth et al., 2004), and biases future research in favour of the erroneous findings (Berlin & Ghersi, 2005).[9][10]

As Berlin and Ghersi describe, a prospective meta-analysis, wherein studies are selected for meta-analysis and pooled together *before* data are produced is the most robust form of meta-analysis, not only preventing publication bias by accessing all data pre-publication but also by enabling analysis of non-aggregated, patient-level data (Berlin & Ghersi, 2005; Clarke & Stewart, 2009; Stewart et al., 2005). Importantly, a prospective meta-analysis provides much greater statistical power than any component study (Berlin & Ghersi, 2005). However, such

[9] #BIASIDENTIFICATION: I explain why bias continues to permeate meta-analyses and discuss how biases at the study level permeate meta-analyses even in the face of mitigation strategies like pre-registration.

[10] #NS125-WORKINCONTEXT: I provide examples of the real-world impacts of misleading conclusions to justify the importance of minimising that bias.

studies are also incredibly resource-intensive, and while they are the ideal, they certainly are not the norm. As such, many clinical trials and meta-analyses of those trials likely will not be prospective, and so we must find other ways to address publication bias.[11]

Detecting and quantifying bias

The Achilles' heel of meta-analysis is that their premise is often faulty: if multiple studies are in concordance, then the effect is likely true (Ioannidis, 2005), but in the face of *p*-hacking, the studies themselves are not reliable (Simonsohn et al., 2014b). Moreover, we saw in the previous section that designing a meta-analysis robust to bias is exceedingly costly and difficult. This motivates a need for retrospective techniques that may detect, quantify and correct for publication bias.

Early tools for measuring this bias include funnel plots (Sterne et al., 2005), regression (Sterne & Egger, 2005), failsafe *N* (Becker, 2005), and trim-and-fill (Duval, 2005). Though each aims to at least detect publication bias, none do so with very much success. Here I will outline each technique and describe its flaws. The cited chapters from Rothstein et al. (2005) offer more detail to the interested reader.[12]

Failsafe N

One of the first measures of bias in meta-analysis, Failsafe *N* stems from Rosenthal's file drawer problem (Rosenthal, 1979) and describes the number of negative or null studies that need to be in the file drawer for the reported effect sizes to be nullified (Becker, 2005). A large *N* means that the effect is robust to publication bias, as it requires a large number of unpublished results, whereas a small *N* suggests strong publication bias, in that only a few studies needed to have been "shelved" for the effect size to be significant.

[11] #BIASMITIGATION: I describe prospective meta-analyses as the gold standard for a bias-mitigated meta-analysis. I then outline the practical limitations of this technique to motivate the need for other statistical methods.

[12] #BIASMITIGATION: In the following sections I describe each statistical bias mitigation technique and explain its flaws.

Though theoretically sound, there is no standard formula for it, and different formulas give vastly different results. Moreover, what constitutes a sufficiently large failsafe *N* is subjective, and subjectivity opens the door to bias (Becker, 2005). The theoretical foundations of Failsafe *N* are also precarious: the statistic does not care about the sample size of the shelved studies, despite observed effect size depending on sample size and the true effect size (Sutton et al., 2000). Though it is a convenient statistic, its assumptions make it a poor tool to reasonably measure publication bias.

Funnel plot

The funnel plot (Figure 2) is a graphical way of detecting the possibility of bias in a body of studies. Its assumption is simple: studies with a greater sample size are more concordant of an effect size than those with smaller sample sizes. When you plot effect size against sample size in a body of research, the non-biased plot will be funnel-shaped (Sterne et al., 2005). Deviations from the ideal funnel plot are indicative of certain types of bias. Publication bias will yield an asymmetric plot, with those non-results from smaller studies excluded, for instance (Sterne & Egger, 2001; Sterne et al., 2005). Unfortunately, publication bias is not the sole cause of funnel plot asymmetry: not only may other types of bias also yield asymmetries, but so too may data irregularities or artefacts of the meta-analytical procedure (Egger et al., 1997). This makes funnel plots poor at specifically detecting publication bias.

Regression

Regression methods aim to quantify the asymmetry of funnel plots. An obvious advantage to these methods is that they systematise assessment of asymmetry; visual assessment of a funnel plot as described in Sterne and Egger (2001) and Sterne et al. (2005) is subjective and says nothing about the magnitude of publication bias. Multiple regression techniques exist, the principal methods being rank correlation (Begg & Mazumdar, 1994), regressing the normal deviate of the effect studied against its precision (Egger et al., 1997), and meta-regression (Sterne & Egger, 2005).

Regression methods have, however, some notable caveats. When the true effect size is

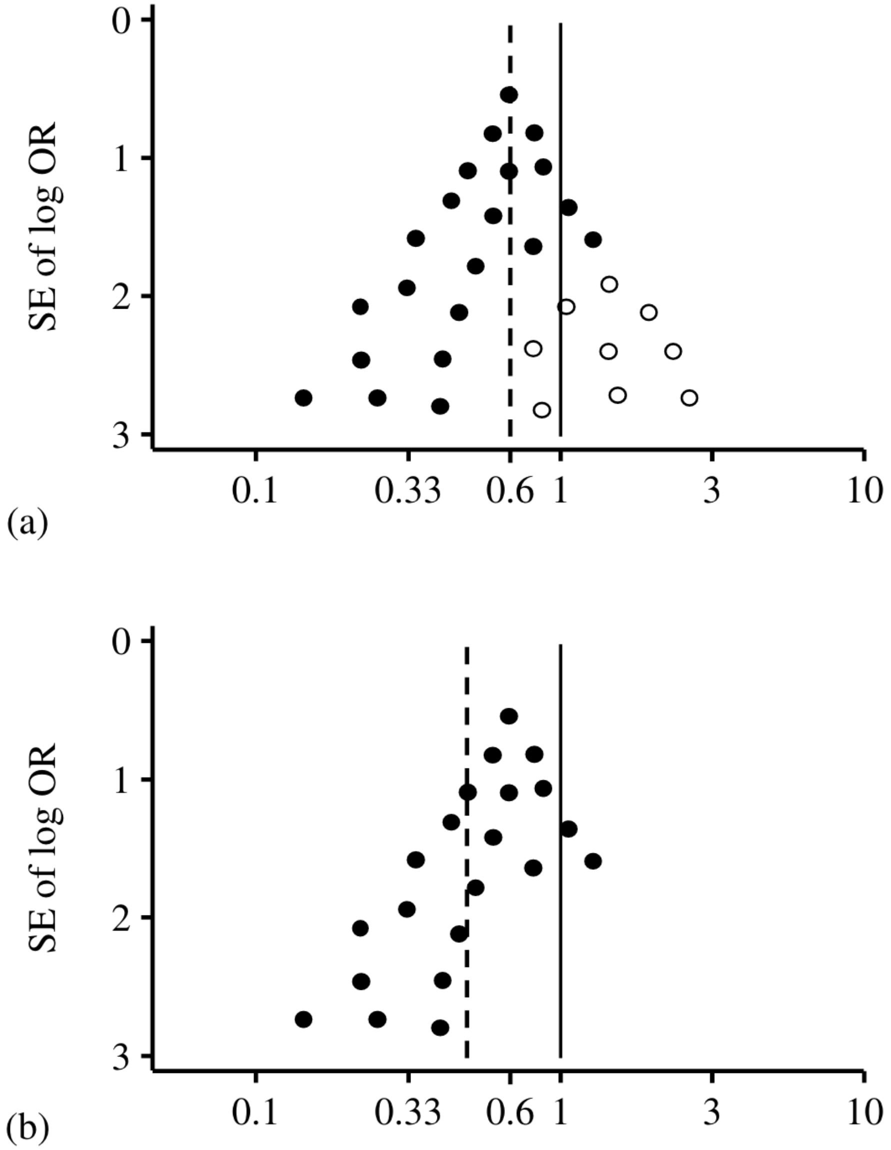

Figure 2
An example of a symmetrical funnel plot representative of a bias-free study sample (top) and an asymmetrical plot wherein small studies showing no effect aren't present. Sterne, J. A., Becker, B. J., & Egger, M. (2005). The Funnel Plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments *(1st ed., pp. 75–98). Wiley. https://doi.org/10.1002/0470870168*

large, when the sample sizes in the studies surveyed are similar, or when a meta-analysis includes fewer than around ten studies, regression methods offer inflated Type I error rate estimates. Moreover, these regression methods inherently assume that sample size predicts publication when it is rather the *p*-value (Sterne & Egger, 2005). Thus regression methods, though they help to remove the subjectivity of a funnel plot assessment, remain a flawed device for measuring *p*-hacking.

Trim-and-fill

Trim-and-fill (Figure 3) differs from the other listed techniques in that it strives to correct, and not just detect, publication bias (Duval, 2005). The basic idea is to make an asymmetric funnel plot symmetric by removing the studies in the asymmetric portion and then using the remaining studies to estimate the centre of the funnel. The removed studies are then placed back onto this adjusted funnel and the effect size is re-estimated. By estimating the funnel's centre without the most biased studies, we adjust the estimated effect size for that bias.

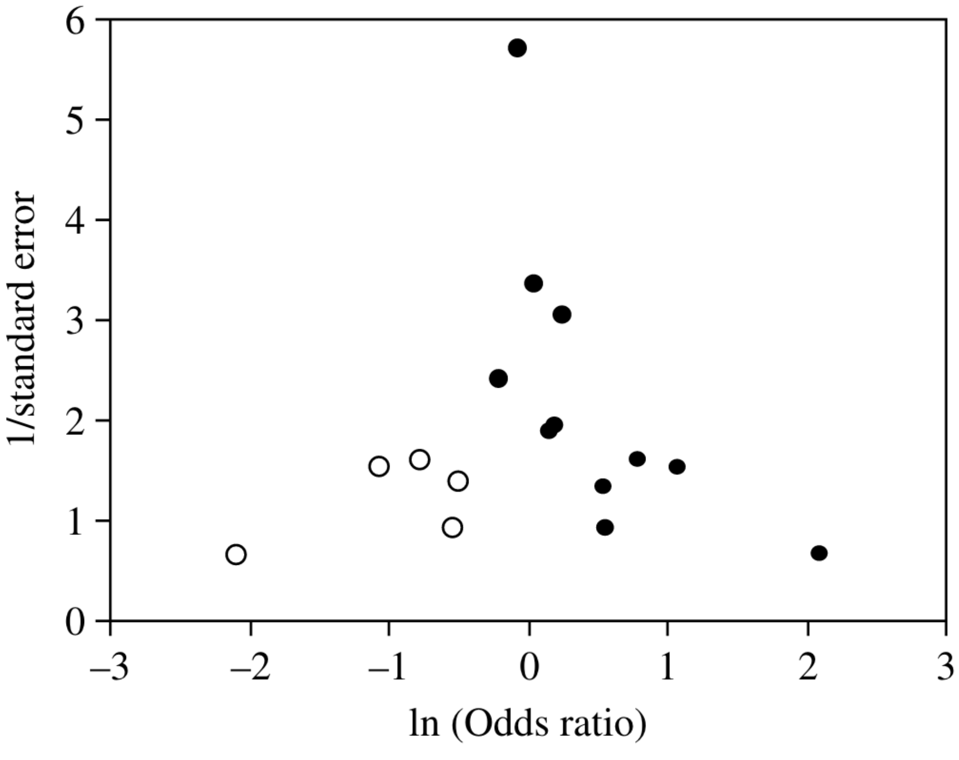

Figure 3
A funnel-plot that has been corrected with trim-and-fill. Unfilled circles are the imputed studies. Duval, S. (2005). The Trim and Fill Method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments *(1st ed., pp. 127–144). Wiley. https://doi.org/10.1002/0470870168*

The technique replaces real biased studies with fictional ones that have been "adjusted" for publication bias. Simulation studies show trim-and-fill to accurately identify missing studies in a meta-analysis (Duval & Tweedie, 2000; Sterne et al., 2000). However, it does so even when bias is not actually present; it succumbs to the same flaw as the funnel plot itself: asymmetries of the funnel arise from more than just publication bias (Egger et al., 1997).

Because trim-and-fill adjusts for bias by replacing missing studies with fictionalised ones, it does not provide accurate estimates of effect sizes. It is thus put forward not as a bias-correcting estimator of effect size, but rather instead as a device for estimating the sensitivity to publication bias in a meta-analysis.

p-Curves

None of the above techniques are totally satisfactory, and particularly lack the ability to correct for upward-skewing biases. The closest to achieving this is the trim-and-fill approach, which, though it does quantify the sensitivity of a meta-analysis to bias, does not provide a mathematically sound approach for then correcting effect sizes with this knowledge.

The *p* curve (Simonsohn et al., 2014b; Simonsohn et al., 2015) alleges to do just that (Figure 4). "*p* curving" has the meta-analyst examine the distribution of significant *p* values reported in the literature. It specifically assesses a research body's 'evidential value' by examining how much selective reporting might explain the findings (Simonsohn et al., 2014b). Much like the funnel plot, the shape of the curve provides information as to the extent and types of bias that permeate the literature (Ulrich & Miller, 2015, 2018). In the absence of bias, the *p* curve is uniform (Ulrich & Miller, 2018). Where there is *p*-hacking (researchers have manipulated findings to be just within the cusp of significance) the curve is left-skewed and where there is publication bias (the most extreme findings are the most likely to be published) the curve is right skewed (Ulrich & Miller, 2018).

While the *p*-curve works rather well, both in detecting and correcting for bias, Ulrich and Miller point out that the opposite skewness under publication bias and *p*-hacking can lead some biased studies to go undetected (2015) and propose an even more robust alternative, the *Z*-curve (Schimmack, 2021). Nonetheless, it is clear from both approaches by Ulrich and Miller and Schimmack that analysing distributions over significance values is a valuable approach (2021,

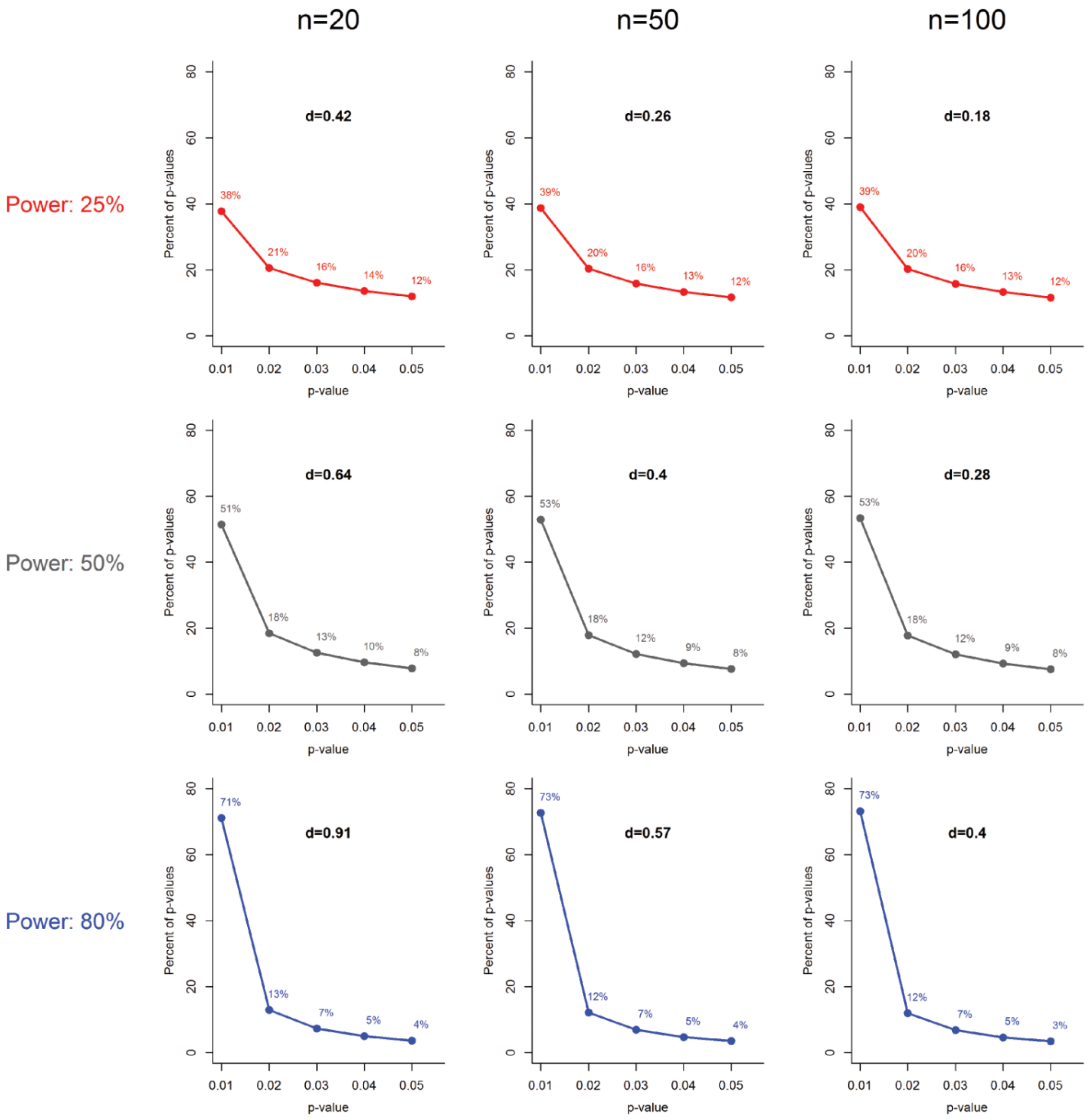

Figure 4

Expected p-curves for different sample sizes and statistical powers. Empirical p-curves can then be compared to the expected p-curve to diagnose bias. Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results [Publisher: SAGE Publications Inc]. Perspectives on Psychological Science, *9(6), 666–681. https://doi.org/10.1177/1745691614553988*

2015)

Returning to Bayes' factors

We saw previously that null-hypothesis testing is rather flawed, and that the modern approach conflates αs and ps and provides researchers with a false sense of comfort that their statistical tests ensure sub-α false positive rates (Hubbard et al., 2003); in the face of statistical manipulation and p-hacking, false-positive rates increase dramatically (Simmons et al., 2011). Bayesian techniques do not have this issue as they do not deal with arbitrary significance thresholds (Ioannidis, 2008). There are no false positives, only measures of evidence and credibility for each hypothesis.[13]

Given the theoretical advantages of Bayesian methods in hypothesis testing, how do they hold up in the face of bias? As a symmetric measure of evidence, B informs us both of the evidence for H_A and H_0. Assuming that both hypotheses are regarded equally well, there ought not to be a need to massage B factors in either direction (Dienes, 2016), though this seems idyllic. Whether their symmetry makes them less prone to bias or not, it remains an advantage of Bayesian methods that they directly interrogate each hypothesis rather than only H_0.

The Bayesian method holds up against other types of statistical manipulation. Selectively adding participants until a significant result is obtained is a surefire route to a false-positive (Simmons et al., 2011). In the Bayesian world, adding participants increases the certainty of the posteriors $P(H_0|D)$ and reduces the weight of the prior (Dienes, 2016). It is, however, possible that adding participants could, by chance, alter B favourably. However, since B is a measure of evidence for or against a model, it would take a very lucky set of participants to alter B to the wrong hypothesis.

Another issue in frequentist testing is that hypothesis tests must be specified beforehand, and that retrospectively designing a test (e.g. choosing one- vs two-tailed) inflates the Type I error rate. Bayesian methods do not have this problem, because they directly assess the probability of the data under each hypothesis; $P(D|H_\alpha)$ is the same whether specified before or after data collection.

However, Bayesian approaches have their limitations. One such limitation is that they do not test hypotheses per se, but rather mathematical models of such hypotheses (Dienes, 2016). How the researcher parameterises and models their theories and mechanisms steers the outcome.

[13] Hypotheses are not accepted or rejected on the basis of B, but rather their *relative likelihood* is quantified.

While this translation from theory to model may encourage researchers to better scrutinise their models (e.g. in psychology) (Dienes, 2016), it also adds to the researcher degrees of freedom, the more of which there are the greater the risk of Type I error (Simmons et al., 2011; Wicherts et al., 2016). Moreover, correctly parameterising a Bayesian model that encapsulates the theory specified by an H_α requires a nuanced understanding of Bayesian statistics, probability distributions and their interpretations and the difficulties of obtaining accurate posteriors, especially with complex models without closed solutions.

While Bayesian approaches avoid many of the pitfalls of NHT, they are also much more complex and mathematically daunting devices that require careful parametrisation. Moreover, they test models of hypotheses, rather than hypotheses themselves (though this is still an improvement over NHT). Though guidance exists on how to parametrise Bayesian hypothesis tests (e.g. Dienes (2021)), they naturally contain vastly more researcher degrees of freedom, from model set-up to model-selection, posterior sampling and model reporting.

Research Questions and Hypotheses

Meta-analyses intend to yield more accurate effect size estimates by providing greater effective sample sizes by pooling together studies that investigate the same effect (Egger et al., 2009; Lipsey & Wilson, 2001) and are certainly less bias-prone than other types of systematic review (Gøtzsche, 1987; Kjaergard & Gluud, 2002; Schmidt & Gøtzsche, 2005). The pervasion of bias, especially publication (Fanelli, 2011; Joober et al., 2012; Pautasso, 2010) and data-dredging (Dickersin et al., 1992; Head et al., 2015; Olson et al., 2002) in the studies that enter the meta-analysis produce biased results despite this (Berlin & Ghersi, 2005; Dickersin, 2005; Simonsohn et al., 2014b; Sutton & Pigott, 2005). Complementary to this issue is a deeply ingrained misunderstanding of hypothesis testing (Hubbard et al., 2003) that threatens the very integrity of the majority of research findings across fields (Ioannidis, 2005).

Bayesian hypothesis testing improves on NHT by directly measuring the evidence for one hypothesis over another (Wagenmakers & Grünwald, 2006) and offers some robustness over certain types of data-dredging (Dienes, 2016), yet also vastly increases the degrees of freedom available to researchers, which Simmons et al. demonstrated to majorly inflate results (2011).

Publication bias and p-hacking, and the implications on meta-analytical estimates and conclusions, are well-understood. While Dienes offer a theoretical basis for the robustness of Bayesian approaches to data-dredging and statistical manipulation (2016), there is little empirical evidence to that point.[14][15]

The goal of this paper is to simulate how bias in individual studies pervades meta-analysis. Each study computes a Bayes' factor B_i per equation 1:

$$B_i = \frac{P(D_i|H_0)}{P(D_i|H_1)} \tag{3}$$

Where D_i is the data from the ith study and we find $P(D|H)$ by integrating over parameter space:

$$P(D_i|H) = \int P(D_i|\theta, H) P(\theta|H) d\theta \tag{4}$$

I will examine bias in a simple difference of means hypothesis test. The null hypothesis is that there is no difference in means, and the alternative is that there is a difference in means. The likelihoods under each hypothesis are then:

$$P(D_i|\theta, H_1) = \mathscr{N}(\mu + \Delta\mu, \sigma = 1)$$
$$P(D_i|\theta, H_0) = \mathscr{N}(\mu, \sigma = 1)$$

With $\mu = 0$. The alternative hypothesis H_1 is two-sided and assumes some difference $\Delta\mu$ between the groups, while the null H_0 does not include the $\Delta\mu$ term and assumes no difference.

[14] #NS125-FRUITFULDIRECTIONS: I concisely summarise the conclusions of the literature review and the scope of our understanding of bias in frequentist NHT. I then describe how that same understanding is significantly weaker in terms of Bayesian hypothesis testing despite its purported advantages. I make note that its robustness to bias as described by Dienes (2016) is purely theoretical and that there is a dearth of empirical evidence to support that claim, thereby motivating the need for further investigation.

[15] #EVIDENCEBASED: I have systematically incorporated a range of empirical studies and scholarly articles to support my analysis of biases in meta-analyses and Bayesian hypothesis testing. I ensure that each claim is substantiated with relevant, high-quality evidence. I have also highlighted the limitations and challenges inherent in these methodologies, acknowledging areas where empirical evidence is lacking, particularly in the context of Bayesian hypothesis testing's resistance to certain biases, thereby motivating the need for my study.

To investigate degrees of bias, I will compare the Bayes' factor computed using different prior distributions that encapsulate varying degrees of bias. Narrower priors make stronger assumptions and are therefore more biased. There are two parameters, μ and $\Delta\mu$. For simplicity, assume that $|\mu| = 1$ and $\sigma = 1$. The prior over μ is then:

$$P(\mu) = \mathscr{U}(-1, 1)$$

We then consider groups each using a different prior over $\Delta\mu$. For analytical simplicity, I use a normal conjugate prior and assume a true effect size of 0.47 (Leucht et al., 2017) (see Methods for justification), and model bias by varying σ: a narrower prior is more biased than a wider one. The priors over $\Delta\mu$ under H_1 that will be compared are then:

$$P_1(\Delta\mu, H_1) = \mathscr{N}(0.47, 5)$$

$$P_2(\Delta\mu, H_1) = \mathscr{N}(0.47, 1)$$

$$P_3(\Delta\mu, H_1) = \mathscr{N}(0.47, 0.5)$$

$$P_4(\Delta\mu, H_1) = \mathscr{N}(0.47, 0.1)$$

$$P_5(\Delta\mu, H_1) = \mathscr{N}(0.47, 0.01)$$

Under H_0, since the likelihood does not depend on $\Delta\mu$, there is no need for a prior over $\Delta\mu$.

The Bayes' factor will be computed for each simulated study per equations 3 and 4. Then, the distributions of the B computed for each study under each prior will be compared.

The difference in Bayes' factors between groups reflects the resilience of the Bayesian test against bias. Another related line of evidence comes from the proportion of B in each group that incorrectly favours H_0. If Bayesian methods are robust to bias, all of the comparison groups will compute similar Bayes' factors, and similar null-hypothesis acceptance rates; the choice of prior will have little impact on the outcome. If, however, they are not robust, we will observe a difference in the Bayes' factors, with groups with narrower, more biased priors tending to produce Bayes' higher Bayes' factors and more strongly favouring H_1 (and a lower proportion of $B < 1$). Given that each study uses a fairly large n, the favouring of H_1, even in the most extreme

$\sigma = 0.01$ group, will still be small.[16]

While the effect of priors within the Bayesian test is important in itself, the primary goal of this investigation is to compare the robustness of the Bayesian method to a frequentist approach. As we're testing a difference of means, a meta-analytical *t*-test presents a logical comparison group. If *B* is indeed a more robust measure than *p*, I expect the false-negative ratio to be lower with the Bayesian test than with the *t*-test for all $\Delta\mu$. Moreover, the change in false negative ratio as $\Delta\mu$ decreases will also be less under the Bayesian paradigm.[17]

Methods

I simulated a simple difference-of-means test with four groups, each with a difference of means $\Delta\mu \in \{0.005, 0.01, 0.05, 0.1, 0.3, 0.5\}$. There are two treatment groups: the placebo O and the treatment T. Then $X_O \sim \mathcal{N}(\mu, \sigma_{X_O} = 1)$ and $X_T \sim \mathcal{N}(\mu + 0.47, \sigma_{X_T} = 1)$ where X is the measured treatment outcome (overall change in symptoms). Each study contains $N = 150$ participants. Assume that each study has equally sized treatment groups.

Each 'study' involves two random draws of $n = 75$ participants from X_O and X_T (Figure 5). Repeating the draws 50 times is akin to a sample of 50 studies. Bayes' factors for each study were then estimated using Equations 3 and 4. The same model was used for both hypotheses, only the distribution of $\Delta\mu$ differs. For H_0:

$$\begin{aligned}
\mu &\sim \mathcal{U}(-1, 1) \\
\Delta\mu &= 0 \\
X_O &\sim \mathcal{N}(\mu, 1) \\
X_T &\sim \mathcal{N}(\mu + \Delta\mu, 1)
\end{aligned}$$

[16] #HYPOTHESISDEVELOPMENT: I present multiple predictions using the "If A then B because C" structure, per feedback from Prof. Wilkins.

[17] #TESTABILITY, #PLAUSIBILITY, #HYPOTHESISDEVELOPMENT, #NS125-QALMRI.

And for H_i:

$$\mu \sim \mathscr{U}(-1,1)$$
$$\Delta\mu_i = P_n(\Delta\mu_i, H_1)$$
$$X_O \sim \mathscr{N}(\mu, 1)$$
$$X_T \sim \mathscr{N}(\mu + \Delta\mu, 1)$$

H_i is the alternative hypothesis for a given $\Delta\mu$ group.

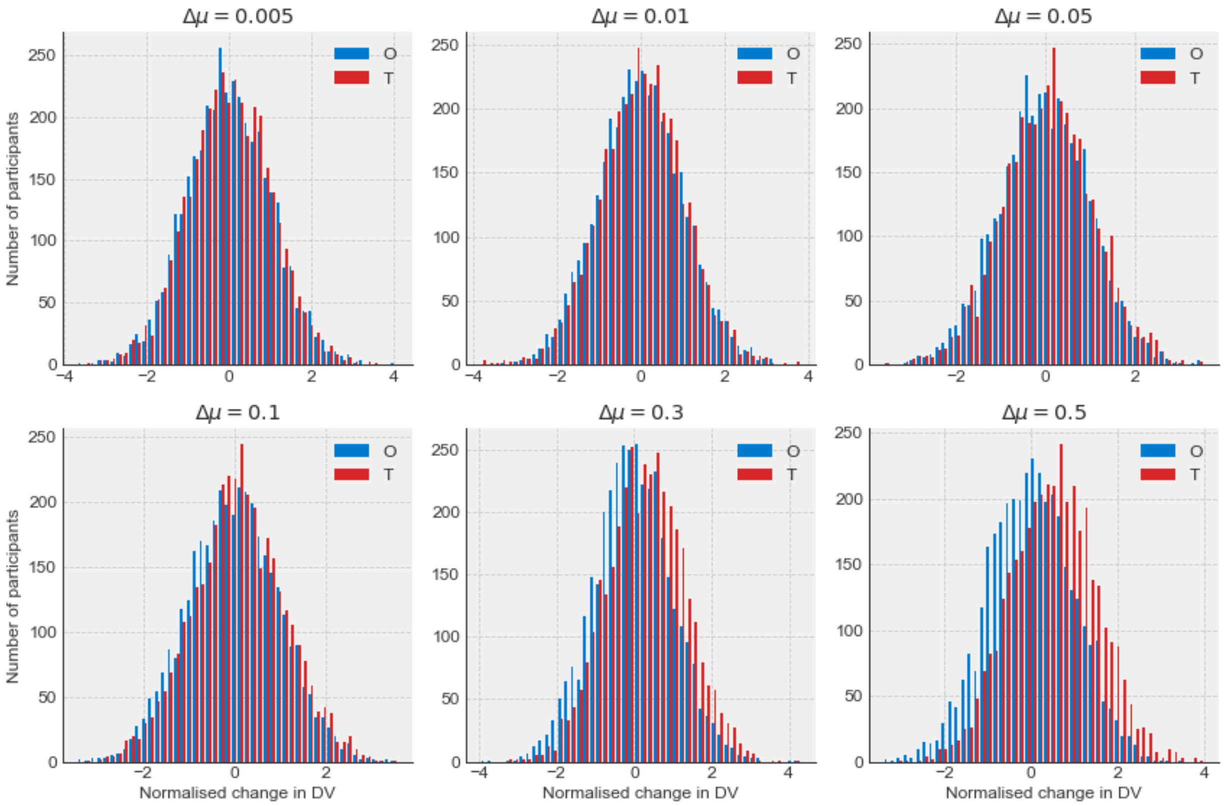

Figure 5

Simulated placebo and treatment data aggregated across all studies with a given $\Delta\mu$

Marginal likelihoods were estimated with a harmonic mean estimator.

The Bayesian models defined by the likelihoods and priors described previously were implemented and sampled using an MCMC NUTS sampler in Pymc. Find the code in the supplemental materials (Appendix A or the attached code notebook).

The simulated data were also assessed with a frequentist t-test. p values for the aggregated data across studies with a given $\Delta\mu$ were computed. A meta-analytical p-value was also computed. Code in Appendix A and the code notebook.[18]

Results

For large $\Delta\mu = 0.5, 0.3$ the majority of the B distribution sits to the right of 1.0; the Bayes' factor favours H_A. For small $\Delta\mu$, however, we see that most of the B distribution is below 1.0; H_0 is favoured. Thus the false negative ratio is greater for small effect sizes, as we'd expect (Figure 8)

By inspection of Figures 6 and 8, the choice of σ_{X_T} impacts the false negative ratio irrespective of $\Delta\mu$. Evident in all plots is that narrower priors favour H_0 more strongly; the proportion of $B < 1$ is greater when σ_{X_T} is smaller. This is a rather surprising result: in frequentist analysis, biased statistical results inflate the false positive ratio. Since wider priors are less conservative as to the existence of a true effect, it makes sense that wider priors incur a higher Type II error rate. Figure 7 supports this, especially the data for $\Delta\mu = 0.5$: $\bar{B}$ is lower for higher σ_{X_T}. When $\Delta\mu$ is smaller, the effect size is too small and the sample size insufficiently small thus why the evidence hovers about $B = 1.0$.

The frequentist meta-analytical difference of means test revealed a non-significant result for $\Delta\mu = 0.005$ and $\Delta\mu = 0.01$; there was a significant difference in means between the placebo and treatment groups for $\Delta\mu \geq 0.05$. Results are presented in Figure 9 and Table 1. This is likely a problem of statistical power: a sample of 150 participants is too small to detect such small effect sizes (there is a true effect since it was encoded into the simulation). A similar problem affected the Bayes' factors.[19]

Discussion

The results suggest that, given a sufficiently large sample size, B is robust to prior hacking. With $\Delta\mu \leq 0.1$, the effect size is too small and the amount of data too little for B to favour H_1,

[18] #NS125-QALMRI: Here I present the methods.

[19] #NS125-QALMRI: Here I present the result in graphical and tabular form, and describe the key findings.

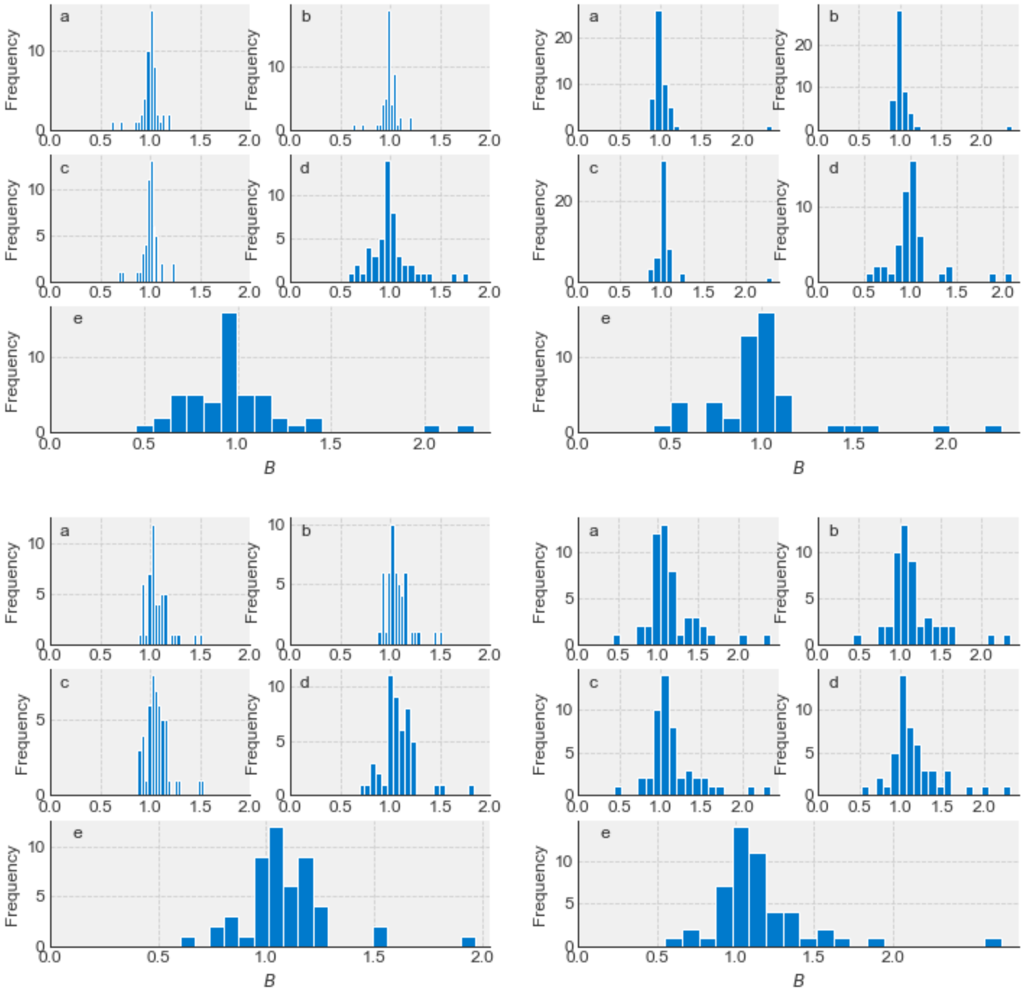

Figure 6

Distribution over $\bar{B}$ for each $\Delta\mu$ and σ_{X_T}. (a) $\sigma_{X_T} = 5$, (b) $\sigma_{X_T} = 1$ (c) $\sigma_{X_T} = 0.5$, (d) $\sigma_{X_T} = 0.1$, (e) $\sigma_{X_T} = 0.01$. Top left: $\Delta\mu = 0.05$, top right: $\Delta\mu = 0.1$, bottom left: $\Delta\mu = 0.3$, bottom right: $\Delta\mu = 0.5$. Stronger evidence for H_1 when $\Delta\mu$ is higher and σ_{X_T} is larger.

thus why B is centred about 1.0 in the upper plots of Figure 7. The distributions of B were fairly consistent irrespective of σ_{X_T} for fixed $\Delta\mu$. This suggests low prior sensitivity: if B *were* sensitive to the prior, we'd expect the distributions on the narrowest priors to have a lower variance than those with higher σ_{X_T}s. Figure 7, plotting the error bars as two standard deviations, has the surprising result of wider posteriors with *narrower* priors, contrary to what we'd expect. This

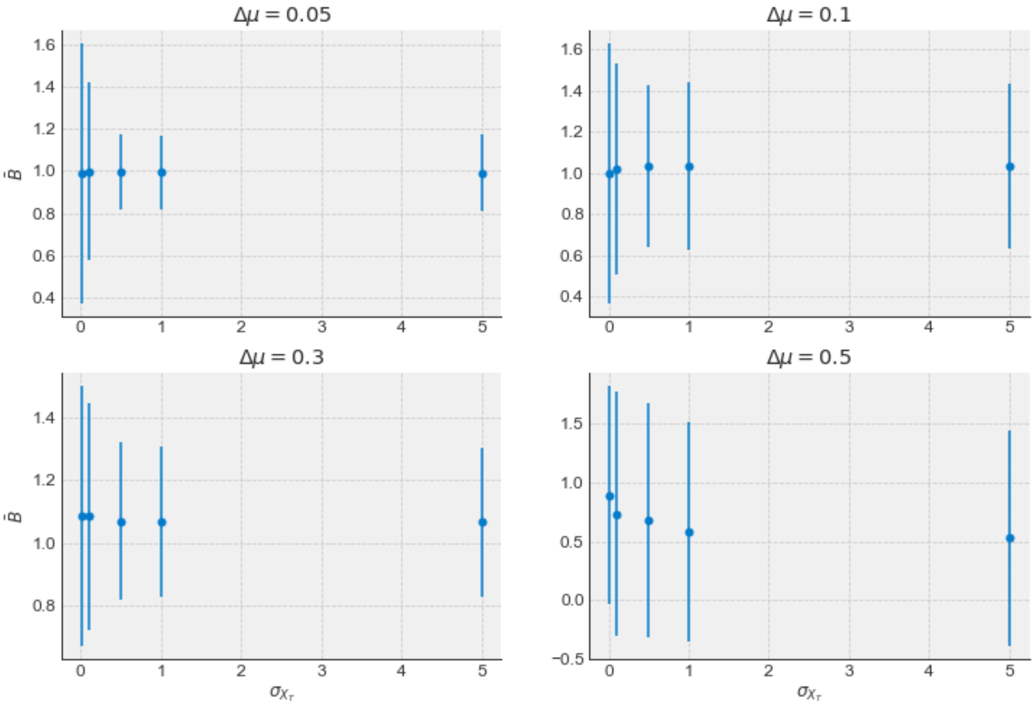

Figure 7

$\bar{B}$ for each σ_{X_T} and $\Delta\mu$. Solid dots represent mean $\bar{B}$ over all studies with a given $\Delta\mu$ and σ_{X_T}. Error bars represent two standard deviations over $\bar{B}$. The $\bar{B}$s are very similar, and the heavily overlapping error bars do not support any impact of σ_{X_T} on B. H_A is slightly favoured with $\Delta\mu = 0.3$ for all σ_T, and very slightly favoured with $\Delta\mu = 0.1$. With $\Delta\mu = 0.05$ there is insufficient evidence for H_1; both hypotheses are equally likely.

$\Delta\mu$	p
0.005	0.86
0.01	0.58
0.05	< 0.01
0.1	< 0.01

Table 1

Meta-analytical p-values for each $\Delta\mu$

demonstrates an interesting property of the Bayesian paradigm, which is that when the prior beliefs are wrong, the likelihood strongly dominates and the net result is less certainty in our beliefs than before.

Across all $\Delta\mu$s, broader priors had a higher Type II error rate (Figure 7). This is because

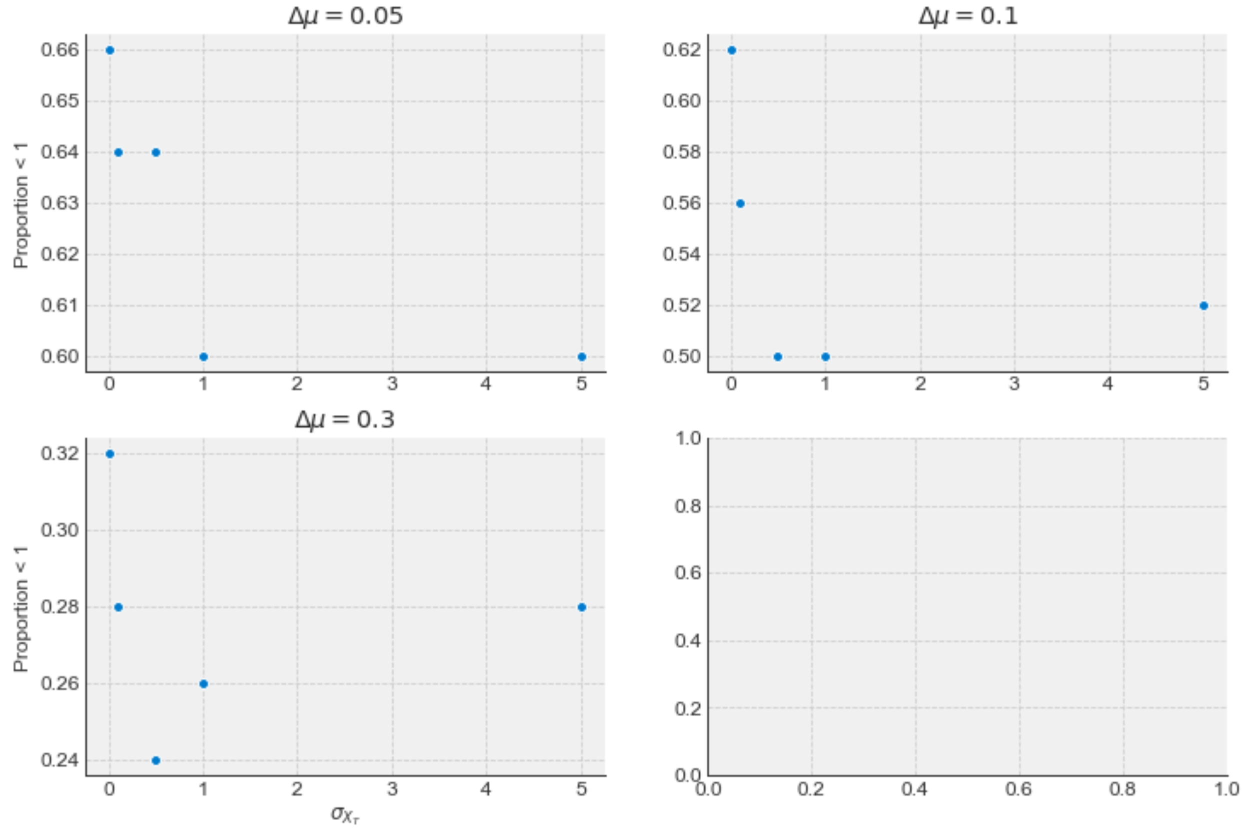

Figure 8

Proportion of B less than one under each σ_{X_T} treatment. Narrower, more biased priors tend to favour H_0 with less frequency than the broader ones.

wider priors are more conservative in favouring the null hypothesis. Such a conservatism is desirable: the Bayesian approach only favours H_1 when the evidence is strongly in favour. Furthermore, this implies that wider priors are less biased. We know this to be qualitatively true by construction, but it is nonetheless reassuring to see the conclusion empirically supported.

As for the false positive rate, we can make some inferences by inspection of data for the $\Delta\mu = 0.05$ group. Though the effect is non-zero, it is, for the given sample size, effectively zero, and the distribution of B is tight around 1.0 for large σ_{X_T} and remains centred around one, though is wider, for small σ_{X_T}. Thus the *probability* of obtaining a Type I error in an individual study is increased by using a biased prior, but, in the overall meta-analysis, we achieve the expected $B = 1.0$.

These preliminary results suggest that the Bayes factor is robust to prior hacking. From Figures 6 and 7, it is clear that the distributions of B across the studies were almost the same,

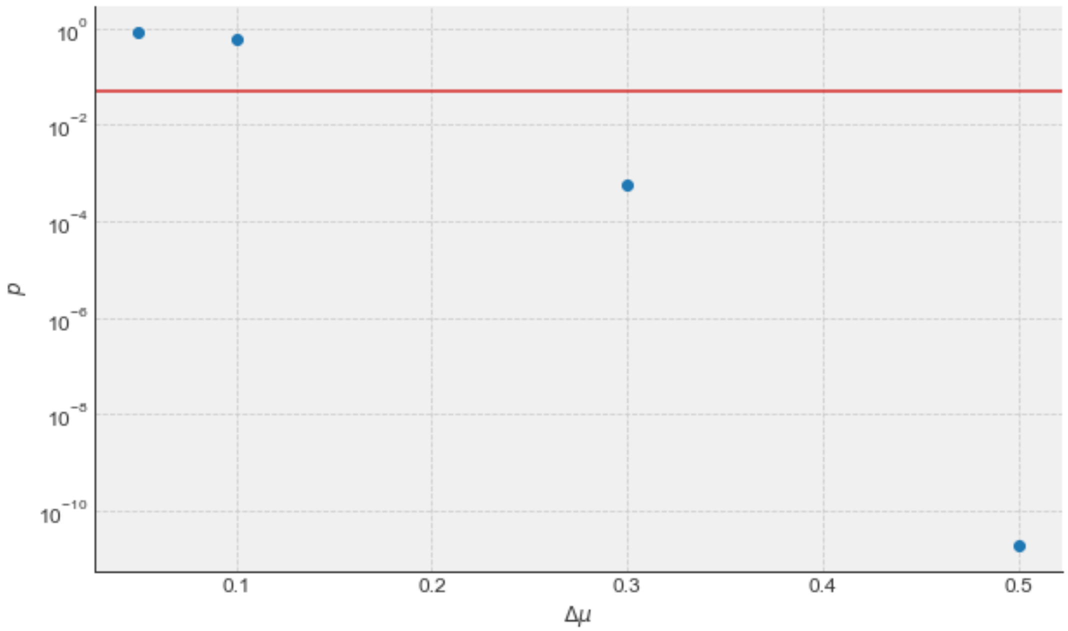

Figure 9

p values for aggregated studies generated with a given $\Delta\mu$. Red line shows $p = 0.05$ significance threshold. Significant difference found for $\Delta\mu \geq 0.05$.

regardless of the choice of σ_{X_T}. This resilience was expected: as each study has a fairly large sample size of $n = 168$ participants, the likelihood function outweighs the prior in determining the posterior.

The evidence suggests that a biased prior decreases the false negative rate in the face of a true result, and increases the false positive rate, both of which are sensical: the data and the researchers' beliefs point in the same direction. When the researchers' bias is contrary to the data, this nets greater uncertainty in the overall evidence (Figure 7), thus a wider prior is always a safer choice, even when there is strong certainty, either from bias or knowledge, in the study's outcome.

Like the Bayesian method, the frequentist approach did not detect a significant difference in means for $\Delta\mu \leq 0.1$. While the Bayesian approach could not favour H_A, the frequentist test had a non-significant p-value. In either case, the conclusion ought to be the same: there is insufficient evidence in the data of there being any effect. This suggests that the sample size limits both the

Bayesian and frequentist approaches, albeit by different mechanisms.[20]

Conclusion

The study demonstrates the robustness of the Bayesian difference-of-means test to bias in prior selection regardless of $\Delta\mu$. Wider priors are more conservative in their favouring of H_1. While this does indeed indicate that Bayesian methods may not be easily manipulated through their priors, their superiority over frequentist approaches remains to be elucidated, at least via this simulation approach. Small effects with $\Delta\mu \leq 0.1$ were not detected by either the frequentist or Bayesian methods. Given their added complexity, both computational and mathematical (compare the code in Appendix A for the meta-analytical p value to the code for the Bayesian model), there is no discernible advantage to the Bayesian method over the frequentist in the context of bias.

Further work needs to be done to investigate the effect of prior manipulation on false positives. Such an investigation could be done by introducing a $\Delta\mu = 0$ group. Furthermore, robustness to bias in other elements of the Bayesian process, like likelihood selection or hypothesis validation by examination of B ought to be further investigated. Furthermore, false positives remain uninvestigated by this study despite the severity of their impacts. Moreover, the study has assumed that all aspects of each component study other than prior selection were perfectly unbiased, yet there are many other degrees of researcher freedom across the data collection process that may impact study outcomes (Simmons et al., 2011), independent of whether the analysis is frequentist or Bayesian. Thus further work in the vein of Simmons et al. (2011) to investigate the robustness of Bayesian methods to biased data collection and processing is necessary.

Given that Bayesian approaches do offer several theoretical and philosophical advantages over frequentist approaches (Hubbard et al., 2003; Ioannidis, 2005; Schervish, 1996; Wagenmakers, 2007; Wagenmakers & Grünwald, 2006), and, though this study has not empirically demonstrated it, they are mathematically less prone to Type I errors (Ioannidis, 2005,

[20] #NS125-QALMRI: Here I interpret the results.

2008), they nonetheless stand as a better alternative to frequentist methods. Work should be done to educate researchers both on the pitfalls of frequentist approaches and Bayesian approaches to move us away from the theoretically flawed frequentist paradigm. The risk of this, however, is that the Bayesian paradigm offers more false-negative, and, in the face of publication bias, more studies in the file drawer.

References

Aghion, P., Dewatripont, M., & Stein, J. C. (2008). Academic freedom, private-sector focus, and the process of innovation. *The RAND Journal of Economics*, *39*(3), 617–635. https://doi.org/10.1111/j.1756-2171.2008.00031.x

Almond, D., Du, X., & Papp, A. (2022). Favourability towards natural gas relates to funding source of university energy centres [Number: 12 Publisher: Nature Publishing Group]. *Nature Climate Change*, *12*(12), 1122–1128. https://doi.org/10.1038/s41558-022-01521-3

Barnes, D. E., & Bero, L. A. (1996). Industry-Funded Research and Conflict of Interest: An Analysis of Research Sponsored by the Tobacco Industry Through the Center for Indoor Air Research. *Journal of Health Politics, Policy and Law*, *21*(3), 515–542. https://doi.org/10.1215/03616878-21-3-515

Bartoš, F., & Schimmack, U. (2022). Z-curve 2.0: Estimating Replication Rates and Discovery Rates. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2021.2720

Becker, B. J. (2005). Failsafe N or File-Drawer Number. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (1st ed., pp. 111–126). Wiley. https://doi.org/10.1002/0470870168

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088–1101.

Begg, C. B., & Berlin, J. A. (1988). Publication Bias: A Problem in Interpreting Medical Data [Publisher: [Wiley, Royal Statistical Society]]. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *151*(3), 419–463. https://doi.org/10.2307/2982993

Berkson, J. (1943). Experience with Tests of Significance: A Reply to Professor R. A. Fisher [Publisher: [American Statistical Association, Taylor & Francis, Ltd.]]. *Journal of the American Statistical Association*, *38*(222), 242–246. https://doi.org/10.2307/2279546

Berlin, J. A., & Ghersi, D. (2005). Preventing Publication Bias: Registries and Prospective Meta-Analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (1st ed., pp. 35–49). Wiley. https://doi.org/10.1002/0470870168

Bolt, H. M. (2011). Publications in toxicology: The current situation. *Archives of Toxicology*, *85*(1), 1–2. https://doi.org/10.1007/s00204-011-0643-z

Callaham, M., Wears, R. L., & Weber, E. (2002). Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA*, *287*(21), 2847–2850. https://doi.org/10.1001/jama.287.21.2847

Christensen-Szalanski, J. J. J., & Beach, L. R. (1984). The citation bias: Fad and fashion in the judgment and decision literature [Place: US Publisher: American Psychological Association]. *American Psychologist*, *39*(1), 75–78. https://doi.org/10.1037/0003-066X.39.1.75

Clarke, M. J., & Stewart, L. (2009). Obtaining individual patient data from randomised controlled trials. In M. Egger (Ed.), *Systematic reviews in health care: Meta-analysis in context* (2. ed., [Nachdr.]). BMJ Books.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed). SAGE Publications.

Davidson, R. A. (1986). Source of funding and outcome of clinical trials. *Journal of General Internal Medicine*, *1*(3), 155–158. https://doi.org/10.1007/BF02602327

Dickersin, K. (1997). How important is publication bias? A synthesis of available data. *AIDS education and prevention: official publication of the International Society for AIDS Education*, *9*(1 Suppl), 15–21.

Dickersin, K., Min, Y. I., & Meinert, C. L. (1992). Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. *JAMA*, *267*(3), 374–378.

Dickersin, K. (2005). Publication Bias: Recognizing the Problem, Understanding Its Origins and Scope, and Preventing Harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis* (pp. 11–35). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470870168.ch2

Dienes, Z. (2016). How Bayes factors change scientific practice [Place: Netherlands Publisher: Elsevier Science]. *Journal of Mathematical Psychology*, *72*, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

Dienes, Z. (2021). How to use and report Bayesian hypothesis tests [Place: US Publisher: Educational Publishing Foundation]. *Psychology of Consciousness: Theory, Research, and Practice*, *8*, 9–26. https://doi.org/10.1037/cns0000258

Duval, S. (2005). The Trim and Fill Method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (1st ed., pp. 127–144). Wiley. https://doi.org/10.1002/0470870168

Duval, S., & Tweedie, R. (2000). A Nonparametric "Trim and Fill" Method of Accounting for Publication Bias in Meta-Analysis [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2000.10473905]. *Journal of the American Statistical Association*, *95*(449), 89–98. https://doi.org/10.1080/01621459.2000.10473905

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, *337*(8746), 867–872. https://doi.org/10.1016/0140-6736(91)90201-Y

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ (Clinical research ed.)*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Egger, M., & Davey Smith, G. (2009). Principles of and procedures for systematic reviews. In M. Egger (Ed.), *Systematic reviews in health care: Meta-analysis in context* (2. ed., [Nachdr.]). BMJ Books.

Egger, M., Davey Smith, G., & O'Rourke, K. (2009). Rationale, potentials, and promise of systematic reviews. In M. Egger (Ed.), *Systematic reviews in health care: Meta-analysis in context* (2. ed., [Nachdr.], pp. 3–19). BMJ Books.

Eisenberg, R. S. (1988). Academic Freedom and Academic Values in Sponsored Research. *Texas Law Review*, *66*.

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries [Publisher: Akadémiai Kiadó, co-published with Springer Science+Business Media B.V., Formerly Kluwer Academic Publishers B.V. Section: Scientometrics]. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Farrell, J. (2016). Corporate funding and ideological polarization about climate change [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *113*(1), 92–97. https://doi.org/10.1073/pnas.1509433112

Finucane, T. E., & Boult, C. E. (2004). Association of funding and findings of pharmaceutical research at a meeting of a medical professional society. *The American Journal of Medicine*, *117*(11), 842–845. https://doi.org/10.1016/j.amjmed.2004.05.029

Fisher, R. A. (1960). Scientific Thought And The Refinement Of Human Reasoning. *Journal of the Operations Research Society of Japan*, *3*(1 & 2). https://orsj.org/wp-content/or-archives50/pdf/e_mag/Vol.03_01_02_001.pdf

Fisher, R. A. (1971). *The Design of Experiments* (9th ed.). Hafner Press.

Fong, D. K. H. (1989). A bayesian approach to the comparison of two means [Publisher: Marcel Dekker, Inc.]. *Communications in Statistics - Theory and Methods*. https://doi.org/10.1080/03610928908830123

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer [Publisher: American Association for the Advancement of Science]. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Freiman, J. A., Chalmers, T. C., Smith, H., & Kuebler, R. R. (1978). The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial: Survey of 71 Negative Trials. *New England Journal of Medicine*, *299*(13), 690–694. https://doi.org/10.1056/NEJM197809282991304

Friese, M., & Frankenbach, J. (2020). P-Hacking and publication bias interact to distort meta-analytic effect size estimates [Place: US Publisher: American Psychological Association]. *Psychological Methods*, *25*(4), 456–471. https://doi.org/10.1037/met0000246

Gøtzsche, P. C. (1987). Reference bias in reports of drug trials. *British Medical Journal (Clinical Research Ed.)*, *295*(6599), 654–656.

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

Hendlin, Y. H., Vora, M., Elias, J., & Ling, P. M. (2019). Financial Conflicts of Interest and Stance on Tobacco Harm Reduction: A Systematic Review [Publisher: American Public Health Association]. *American Journal of Public Health*, *109*(7), e1–e8. https://doi.org/10.2105/AJPH.2019.305106

Hubbard, R., Bayarri, M. J., Berk, K. N., & Carlton, M. A. (2003). Confusion over Measures of Evidence (p's) versus Errors ('s) in Classical Statistical Testing. *The American Statistician*, *57*(3), 171–182. http://www.jstor.org/stable/30037265

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False [Publisher: Public Library of Science]. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2006). Concentration of the Most-Cited Papers in the Scientific Literature: Analysis of Journal Ecosystems [Publisher: Public Library of Science]. *PLOS ONE*, *1*(1), e5. https://doi.org/10.1371/journal.pone.0000005

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology (Cambridge, Mass.)*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Johnson, R. T., & Dickersin, K. (2007). Publication bias against negative results from clinical trials: Three of the seven deadly sins. *Nature Clinical Practice Neurology*, *3*(11), 590–591. https://doi.org/10.1038/ncpneuro0618

Joober, R., Schmitz, N., Annable, L., & Boksa, P. (2012). Publication bias: What are the challenges and can they be overcome? [Publisher: Journal of Psychiatry and Neuroscience Section: Editorial]. *Journal of Psychiatry and Neuroscience*, *37*(3), 149–152. https://doi.org/10.1503/jpn.120065

Kelley, K., & Preacher, K. J. (2012). On effect size [Place: US Publisher: American Psychological Association]. *Psychological Methods*, *17*(2), 137–152. https://doi.org/10.1037/a0028086

Kjaergard, L. L., & Gluud, C. (2002). Citation bias of hepato-biliary randomized clinical trials. *Journal of Clinical Epidemiology*, *55*(4), 407–410. https://doi.org/10.1016/s0895-4356(01)00513-3

Koren, G., & Klein, N. (1991). Bias against negative studies in newspaper reports of medical research. *JAMA*, *266*(13), 1824–1826.

Krimsky, S. (2013). Do Financial Conflicts of Interest Bias Research?: An Inquiry into the "Funding Effect" Hypothesis [Publisher: SAGE Publications Inc]. *Science, Technology, & Human Values*, *38*(4), 566–587. https://doi.org/10.1177/0162243912456271

Leucht, S., Leucht, C., Huhn, M., Chaimani, A., Mavridis, D., Helfer, B., Samara, M., Rabaioli, M., Bächer, S., Cipriani, A., Geddes, J. R., Salanti, G., & Davis, J. M. (2017). Sixty Years of Placebo-Controlled Antipsychotic Drug Trials in Acute Schizophrenia: Systematic Review, Bayesian Meta-Analysis, and Meta-Regression of Efficacy Predictors [Publisher: American Psychiatric AssociationArlington, VA]. *American Journal of Psychiatry*. https://doi.org/10.1176/appi.ajp.2017.16121358

Lexchin, J., Bero, L. A., Djulbegovic, B., & Clark, O. (2003). Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. *BMJ (Clinical research ed.)*, *326*(7400), 1167–1170. https://doi.org/10.1136/bmj.326.7400.1167

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* [Pages: ix, 247]. Sage Publications, Inc.

Malone, R. E. (2013). Changing Tobacco Control's policy on tobacco industry-funded research [Publisher: BMJ Publishing Group Ltd Section: Editorial]. *Tobacco Control*, *22*(1), 1–2. https://doi.org/10.1136/tobaccocontrol-2012-050874

Manzoli, L., Flacco, M. E., D'Addario, M., Capasso, L., De Vito, C., Marzuillo, C., Villari, P., & Ioannidis, J. P. A. (2014). Non-publication and delayed publication of randomized trials on vaccines: Survey. *BMJ*, *348*(may16 1), g3058–g3058. https://doi.org/10.1136/bmj.g3058

McShane, B. B., & Gelman, A. (2022). Selecting on statistical significance and practical importance is wrong. *Journal of Information Technology*, *37*(3), 312–315. https://doi.org/10.1177/02683962221086297

Michaels, P. J. (2008). Evidence for "Publication Bias" concerning Global Warming in Science and Nature [Publisher: SAGE Publications Ltd STM]. *Energy & Environment*, *19*(2), 287–301. https://doi.org/10.1260/095830508783900735

Neyman, J., & Pearson, E. S. (1928). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I [Publisher: [Oxford University Press, Biometrika Trust]]. *Biometrika*, *20A*(1/2), 175–240. https://doi.org/10.2307/2331945

Olson, C. M., Rennie, D., Cook, D., Dickersin, K., Flanagin, A., Hogan, J. W., Zhu, Q., Reiling, J., & Pace, B. (2002). Publication Bias in Editorial Decision Making. *JAMA*, *287*(21), 2825–2828. https://doi.org/10.1001/jama.287.21.2825

Pautasso, M. (2010). Worsening file-drawer problem in the abstracts of natural, medical and social science databases. *Scientometrics*, *85*(1), 193–202. https://doi.org/10.1007/s11192-010-0233-5

Procyshyn, R. M., Chau, A., Fortin, P., & Jenkins, W. (2004). Prevalence and outcomes of pharmaceutical industry-sponsored clinical trials involving clozapine, risperidone, or olanzapine. *Canadian Journal of Psychiatry. Revue Canadienne De Psychiatrie*, *49*(9), 601–606. https://doi.org/10.1177/070674370404900905

Rising, K., Bacchetti, P., & Bero, L. (2008). Reporting Bias in Drug Trials Submitted to the Food and Drug Administration: Review of Publication and Presentation (J. Ioannidis, Ed.). *PLoS Medicine*, *5*(11), e217. https://doi.org/10.1371/journal.pmed.0050217

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Roth, D., Boyle, E., Beer, D., Malik, A., & deBruyn, J. (2004). Depressing research. *Lancet (London, England)*, *363*(9426), 2087. https://doi.org/10.1016/S0140-6736(04)16462-3

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication Bias in Meta-Analysis*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470870168.ch2

Schervish, M. J. (1996). P Values: What They are and What They are Not [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.1996.10474380]. *The American Statistician*, *50*(3), 203–206. https://doi.org/10.1080/00031305.1996.10474380

Schimmack, U. (2021). Z-Curve: An even better p-curve. Retrieved April 8, 2023, from https://replicationindex.com/2021/04/25/z-curve-an-even-better-p-curve/

Schmidt, L. M., & Gøtzsche, P. C. (2005). Of mites and men: Reference bias in narrative review articles [Publisher: Frontline Medical Communications]. *Journal of Family Practice*, *54*(4), 334–338. Retrieved August 6, 2023, from https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=16697440&site=ehost-live

Scott, A., Rucklidge, J. J., & Mulder, R. T. (2015). Is Mandatory Prospective Trial Registration Working to Prevent Publication of Unregistered Trials and Selective Outcome Reporting? An Observational Study of Five Psychiatry Journals That Mandate Prospective Clinical Trial Registration (J. M. Wicherts, Ed.). *PLOS ONE*, *10*(8), e0133718. https://doi.org/10.1371/journal.pone.0133718

Seward, D. M. (1999). Criteria for manuscript assessment of scientific dental journals — a postal survey [Number: 7 Publisher: Nature Publishing Group]. *British Dental Journal*, *187*(7), 374–374. https://doi.org/10.1038/sj.bdj.4800283a1

Simes, R. J. (1986). Publication bias: The case for an international registry of clinical trials. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *4*(10), 1529–1541. https://doi.org/10.1200/JCO.1986.4.10.1529

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results [Publisher: SAGE Publications Inc]. *Perspectives on Psychological Science*, *9*(6), 666–681. https://doi.org/10.1177/1745691614553988

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

Sterne, J. A., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, *54*(10), 1046–1055. https://doi.org/10.1016/s0895-4356(01)00377-8

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53*(11), 1119–1129. https://doi.org/10.1016/s0895-4356(00)00242-0

Sterne, J. A., Becker, B. J., & Egger, M. (2005). The Funnel Plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (1st ed., pp. 75–98). Wiley. https://doi.org/10.1002/0470870168

Sterne, J. A., & Egger, M. (2005). Regression Methods to Detect Publication and Other Bias in Meta-Analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (1st ed., pp. 99–110). Wiley. https://doi.org/10.1002/0470870168

Stewart, L., Tierney, J., & Burdett, S. (2005). Do Systematic Reviews Based on Individual Patient Data Offer a Means of Circumventing Biases Associated with Trial Publications? In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (1st ed., pp. 261–286). Wiley. https://doi.org/10.1002/0470870168

Sutton, A. J., Song, F., Gilbody, S. M., & Abrams, K. R. (2000). Modelling publication bias in meta-analysis: A review. *Statistical Methods in Medical Research*, *9*(5), 421–445. https://doi.org/10.1177/096228020000900503

Sutton, A. J. (2005). Evidence Concerning the Consequences of Publication and Related Biases. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (1st ed.). Wiley. https://doi.org/10.1002/0470870168

Sutton, A. J., & Pigott, T. D. (2005). Bias in Meta-Analysis Induced by Incompletely Reported Studies. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (1st ed.). Wiley. https://doi.org/10.1002/0470870168

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology*, *53*(2), 207–216. https://doi.org/10.1016/S0895-4356(99)00161-4

Ulrich, R., & Miller, J. (2015). P-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General*, *144*(6), 1137–1145. https://doi.org/10.1037/xge0000086

Ulrich, R., & Miller, J. (2018). Some properties of p-curves, with an application to gradual publication bias [Place: US Publisher: American Psychological Association]. *Psychological Methods*, *23*, 546–560. https://doi.org/10.1037/met0000125

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., & Grünwald, P. (2006). A Bayesian Perspective on Hypothesis Testing.

Wei, Z., Yang, A., Rocha, L., Miranda, M. F., & Nathoo, F. S. (2022). A Review of Bayesian Hypothesis Testing and Its Practical Implementations [Number: 2 Publisher: Multidisciplinary Digital Publishing Institute]. *Entropy*, *24*(2), 161. https://doi.org/10.3390/e24020161

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in*

Psychology, *7*. Retrieved October 25, 2023, from https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01832

Yaphe, J., Edman, R., Knishkowy, B., & Herman, J. (2001). The association between funding by commercial interests and study outcome in randomized controlled drug trials. *Family Practice*, *18*(6), 565–568. https://doi.org/10.1093/fampra/18.6.565

Zou, C. X., Becker, J. E., Phillips, A. T., Garritano, J. M., Krumholz, H. M., Miller, J. E., & Ross, J. S. (2018). Registration, results reporting, and publication bias of clinical trials supporting FDA approval of neuropsychiatric drugs before and after FDAAA: A retrospective cohort study. *Trials*, *19*(1), 581. https://doi.org/10.1186/s13063-018-2957-0

Appendix A

Supplemental materials

Code

Simulating studies

```python
#in this cell I simulate the studies
import numpy as np
from scipy import stats as sts
from matplotlib import pyplot as plt

Num_studies = 50 #the number of studies
N = 150 #number of participants

#to get the mean of the placebo group I've taken the mean
mu = 0 #mean of placebo group. The meta-analysis used a standardised mean effect size
#so the placebo group is just a standard normal
sigma = 1 #standard deviation of treatment group
Deltamus = [0.005, 0.01, 0.05, 0.1, 0.3, 0.5] #difference between groups


def study(N,mu,sigma, Deltamu): #this funtion runs a single study
    O = sts.norm.rvs(loc = mu, scale = sigma, size = N//2) #generate N/2 samples from the placebo dist
    T = sts.norm.rvs(loc = mu+Deltamu, scale = sigma, size = N//2) #generate N/2 samples from the treatment dist
    return O, T

def run_studies(Num_studies, N, mu, sigma, Deltamu):
    studies_O = []
    studies_T = []
    for i in range(Num_studies):
        O, T = study(N, mu, sigma, Deltamu)  # run the study function and unpack the result
        studies_O.append(O)  # append the first result to studies_O
        studies_T.append(T)  # append the second result to studies_T
    return studies_O, studies_T


studies_O_all = [] #list of lists of all of the studies for each Deltamu, placebo
studies_T_all = [] #list of lists of all of the studies for each Deltamu, treatment

#studies_O, studies_T = run_studies(Num_studies, N, mu, sigma, Deltamu)
#need to separately do O and T
def plot_agg_study_results(studies_O, studies_T, ax):
    agg_results = [np.concatenate(studies_O), np.concatenate(studies_T)]
    ax.hist(agg_results, bins=50, color=color_palette[0:2])
    ax.legend([’O’,’T’])

def plot_by_Deltamu(Num_studies, N, mu, sigma, Deltamus, studies_O_all, studies_T_all):
    fig, axs = plt.subplots(2, 3, figsize=(12, 8))
    axs = axs.flatten()  # flatten to easily iterate over
    for i, Deltamu in enumerate(Deltamus):
        plot_agg_study_results(studies_O_all[i], studies_T_all[i], axs[i])
        axs[i].set_title(f’$\Delta\mu = {Deltamu}$’)

        # conditionally set x and y labels
        if i >= 3:  # bottom row subplots
            axs[i].set_xlabel(’Normalised change in DV’)
        if i % 3 == 0:  # left column subplots
            axs[i].set_ylabel(’Number of participants’)

    plt.tight_layout()
    plt.savefig(’agg_data_hist.png’)  # Move savefig out of plot_agg_study_results

def run_by_delta_mu(Num_studies, N, mu, sigma, Deltamus):
    #runs Num_studies studies with N participants for each Deltamu.
    studies_O_all = []
    studies_T_all = []
    for i, Deltamu in enumerate(Deltamus):
        #generating data for a given Deltamu
        studies = run_studies(Num_studies, N, mu, sigma, Deltamu)
        studies_0_all.append(studies[0]) #append the placebo
        studies_T_all.append(studies[1]) #append the treatment
    return studies_0_all, studies_T_all #returns a list of lists

#generating all of the data
studies_0_all, studies_T_all = run_by_delta_mu(Num_studies, N, mu, sigma, Deltamus)

plot_by_Deltamu(Num_studies, N, mu, sigma, Deltamus, studies_0_all, studies_T_all)
```

Calculating Bayes’ factors

```
#In this cell I run the null and alternative models and then estimate the bayes’ factor for each.
import pymc as pm
n_samples = 2000
#null model
def bayes_model_m0(study_0, study_T):
    with pm.Model() as m0:
        mu = pm.Uniform(’mu’,-1,1) #prior
        delta_mu = 0 #under the null model delta mu is 0
        x_0 = pm.Normal(’x_0’, mu = mu, sigma = 1, observed = study_0) #placebo
        x_T = pm.Normal(’x_T’, mu = mu+delta_mu, sigma = 1, observed = study_T)#treatment

        #sample the model
        trace_m0 = pm.sample(n_samples, idata_kwargs = {’log_likelihood’:True})

    return trace_m0
#alternative model. Includes parameter sigma_T which alters the width of the prior over delta mu
def bayes_model_m1(study_0, study_T, sigma_T):
    with pm.Model() as m1:
        mu = pm.Uniform(’mu’,-1,1) #prior
        delta_mu = pm.Normal(’delta_mu’, mu = 0.47, sigma = sigma_T) #delta mu is distributed
        x_O = pm.Normal(’x_O’, mu = mu, sigma = 1, observed = study_O) #placebo
        x_T = pm.Normal(’x_T’, mu = mu+delta_mu, sigma = 1, observed = study_T) #treatment

        #sample the model
        trace_m1 = pm.sample(n_samples, idata_kwargs = {’log_likelihood’:True})
    return trace_m1
#estimates the marginal likelihood of x_T
def harmonic_mean_estimator(trace):
    log_likelihoods = trace.log_likelihood.x_T.values.flatten() #extracting log likelihoods

    # Convert log-likelihoods to likelihoods
    likelihoods = np.exp(log_likelihoods) #undoing the log

    # Calculate the harmonic mean of the likelihoods
    harmonic_mean = len(likelihoods) / np.sum(1.0 / likelihoods) #computing harmonic mean

    return harmonic_mean

#computes the bayes factor given alternative and null marginal likelihoods
def bayes_factor(ml0,ml1):
    return ml1/ml0

#computes Bayes’ factors for each study in a sample of studies with one alternative and one null model
def simulate_all_studies(Num_studies, sigma_T):
    B_list = []
    for i in range(Num_studies):
        trace_m0 = bayes_model_m0(studies_O[i], studies_T[i])
        trace_m1 = bayes_model_m1(studies_O[i], studies_T[i], sigma_T)

        ml0 = harmonic_mean_estimator(trace_m0)
        ml1 = harmonic_mean_estimator(trace_m1)

        B = bayes_factor(ml0,ml1)
        B_list.append(B)
    return B_list

#repeats the simulation over different alternative priors
def simulate_over_sigma_T(N_studies, sigma_Ts):
    B_all = []

    for sigma_T in sigma_Ts:
        B_list = simulate_all_studies(N_studies, sigma_T)
        B_all.append(B_list)

    return B_all

sigma_Ts = [5,1,0.5,0.1,0.01] #the standard deviations for the alternate priors over Deltamu
data = simulate_over_sigma_T(50, sigma_Ts)
```

Generating plots

```
#this cell generates plots of the simulated Bayes’ factors

#histograms of Bayes’ factors given n comparison groups
import matplotlib.gridspec as gridspec
def plot_b_factors_2(data, sigma_Ts, parent_ax, color_palette):
    n = len(data)
    subplot_labels = [’a’, ’b’, ’c’, ’d’, ’e’]

    if n == 5:
```

```
    nrows = 3
    ncols = 2
    nested_gs = gridspec.GridSpecFromSubplotSpec(nrows, ncols, subplot_spec=parent_ax)

    for i in range(4):
        ax = plt.subplot(nested_gs[i//2, i%2])
        ax.hist(data[i], bins=20, edgecolor=’white’, color=color_palette[0])
        ax.set_xlabel(’$B$’)
        ax.set_ylabel(’Frequency’)
        ax.text(0.05, 0.95, subplot_labels[i], transform=ax.transAxes, fontsize=12, va=’top’, ha=’left’)
        ax.set_xticks(np.linspace(0, 2, num=5))
        ax.axvline(np.mean(data[i]), color=color_palette[1], linestyle=’--’)

    ax = plt.subplot(nested_gs[2, :])
    ax.hist(data[4], bins=20, edgecolor=’white’, color=color_palette[0])
    ax.set_xlabel(’$B$’)
    ax.set_ylabel(’Frequency’)
    ax.text(0.05, 0.95, subplot_labels[4], transform=ax.transAxes, fontsize=12, va=’top’, ha=’left’)
    ax.set_xticks(np.linspace(0, 2, num=5))
    ax.axvline(np.mean(data[4]), color=color_palette[1], linestyle=’--’)
else:
    nrows = n // 2 + n % 2
    ncols = 2 if n > 1 else 1
    nested_gs = gridspec.GridSpecFromSubplotSpec(nrows, ncols, subplot_spec=parent_ax)

    for i in range(n):
        ax = plt.subplot(nested_gs[i])
        ax.hist(data[i], bins=20, edgecolor=’white’, color=color_palette[0])
        ax.set_xlabel(’$B$’)
        ax.set_ylabel(’Frequency’)
```

```
            ax.text(0.05, 0.95, subplot_labels[i], transform=ax.transAxes, fontsize=12, va=’top’, ha=’left’)
            ax.set_xticks(np.linspace(0, 2, num=5))
            ax.axvline(np.mean(data[i]), color=color_palette[1], linestyle=’--’)

fig = plt.figure(figsize=(12, 12))
gs = gridspec.GridSpec(2, 2, figure=fig)

for i, gs_sub in enumerate(gs):
    # Create a subplot in the main figure for each of the 4 sections
    parent_ax = fig.add_subplot(gs_sub)
    parent_ax.set_xticks([])
    parent_ax.set_yticks([])
    parent_ax.spines[’top’].set_visible(False)
    parent_ax.spines[’right’].set_visible(False)
    parent_ax.spines[’bottom’].set_visible(False)
    parent_ax.spines[’left’].set_visible(False)

    # Call the plot_b_factors function for each set of data, passing the appropriate subplot
    plot_b_factors_2(all_data[i], sigma_Ts, gs_sub, color_palette)



plt.savefig(’nested_bayes_hist.png’)
plt.show()

Dmus = [0.05, 0.1, 0.3, 0.5]

all_data = [data_005, data_01, data_03, data]

def plot_B_mean_Dmus(data, Dmus):
    fig, axs = plt.subplots(2,2, figsize = (12,8))
    axs = axs.flatten()
    for i, Dmu in enumerate(Dmus):
        plot_B_mean(data[i], axs[i])
        axs[i].set_title(f’$\Delta\mu = {Dmu} $’)

        #conditionally setting axis labels
        if i>= 2: #bottom row subplots
            axs[i].set_xlabel(’$\sigma_{X_T}$’)

        if i % 2 == 0:
            axs[i].set_ylabel(r’$\bar{B}$’)

    plt.savefig(’b_bar_subplots.png’)  # Saving the figure containing all subplots
    plt.show()

def plot_proportion_less_than_one_Dmus(data, sigma_Ts, Dmus):
    fig, axs = plt.subplots(2,2, figsize = (12,8))
    axs = axs.flatten()
    for i, Dmu in enumerate(Dmus):
        plot_proportion_less_than_one(data[i], sigma_Ts, axs[i])
        axs[i].set_title(f’$\Delta\mu = {Dmu} $’)

        #conditionally setting axis labels
        if i>= 2: #bottom row subplots
            axs[i].set_xlabel(’$\sigma_{X_T}$’)

        if i % 2 == 0:
            axs[i].set_ylabel(’Proportion < 1’)
    plt.savefig(’proportion_less_than_one_subplots.png’)
    plt.show()
```

Meta-analytical power

```
def meta_analytical_p(Dmu):
    es = Dmu  # Enter your summary effect size
    as_ = 50   # Average per number per group (renamed to as_ to avoid keyword conflict)
    mk = 50    # Number of effect sizes
    hg = 3     # Heterogeneity (".33" for small, "1" for moderate, & "3" for large)

    eq1 = ((as_ + as_) / ((as_) * (as_))) + ((es) / (2 * (as_ + as_)))
    eq2 = hg * (eq1)
    eq3 = eq2 + eq1
    eq4 = eq3 / mk
    eq5 = (es / (eq4 ** 0.5))
    Power = (1 - norm.cdf(1.96 - eq5))  # Two-tailed
    result = 1 - Power
    return result

p_values = []
for Dmu in Dmus:
    p_values.append(meta_analytical_p(Dmu))

def plot_p_v_deltamu(Deltamus, p_values):
    #creates a plot of the p value vs delta mu
    plt.figure(figsize=(10, 6))
    plt.scatter(Deltamus, p_values)
    plt.axhline(0.05, color =’#d62728’ )
    plt.xlabel(’$\Delta\mu$’)
    plt.ylabel(’$p$’)
    plt.yscale(’log’)  # using logarithmic scale for better visibility of small p-values
    plt.grid(True, which="both", ls="--")
    plt.savefig(’pvdmu.png’)
    plt.show()

plot_p_v_deltamu(Dmus, p_values)
```

Appendix B

AI Integrity Statement

I used ChatGPT to guide, assist and offer feedback throughout the Capstone process. I first used ChatGPT 4.0 to provide an initial, week-by-week breakdown of the project as a starting point that I then adapted as I went. You can view the conversation here. The plan made my ChatGPT is far off from the one I ended up following, but its initial breakdown of the steps provided a useful framework when I went to build my own deadlines.

I also used ChatGPT to assist with research and editing. Using the "ScholarAI" and "AIPDF" plug-ins, ChatGPT helped me to identify sources on a few topics and summarise academic articles. This was helpful for screening whether an article would be useful for my literature review and determining if it would be worth my while to read and eventually cite. ChatGPT summarised the following bibliography entries for me: Berlin and Ghersi (2005) and Dienes (2016). View the conversations here and here. ChatGPT identified and summarised the following studies that were cited: Manzoli et al. (2014), Rising et al. (2008), and Zou et al. (2018). You may view the conversation here. After reading the summaries provided by ChatGPT, I then read the articles with 2/3 passes ("three pass approach") to verify the conclusions made by ChatGPT before citing these studies in my work.

I used ChatGPT iteratively to gain feedback on my implementation of HCs. Here is an example of a conversation where I asked ChatGPT for feedback on #BIASIDENTIFICATION in the section "*p*-hacking." My approach was to provide ChatGPT with the common questions and pitfalls sections on the HC from the HC Handbook and to then provide it with the section in which I believed I applied the HC. My custom instructions ask ChatGPT to provide honest feedback as I have noticed it to be a bit of a flatterer. Though ChatGPT's feedback is no substitute for my own assessment of my work, and especially not for feedback from my graders, it was nonetheless helpful in identifying ways I could improve my HC applications.

Appendix C

HCs & LOs

#NAVIGATION:

I began work in May 2023, reading articles about hypothesis testing and summarising my findings in a Google Doc. I set the ambitious goal of finishing the literature review by the beginning of the Fall semester. Though I did not quite achieve this, it meant that I could focus the latter half of the Fall semester on a prototype of the full study. Per Prof Wilkins' advice, my goal was to have a full working draft of the Capstone's major components by the December deadline. By writing the simulation with compartmentalised code, expanding my work to the full investigation presented here was straightforward in terms of coding, and left plenty of time for refining my HC/LO applications and prose. My goal for the Spring was to present a Capstone of final deadline quality to the Revised Draft deadline. Thus I was left with only minor HC/LO improvements (and a buggy code cell) leading up to the final submission, making this a relatively stress-free last lap. Moreover, it allowed plenty of time to copy edit and refine the body across multiple sittings; nothing presented in this paper is a first draft. By setting ambitious goals for each deadline, I effectively worked ahead: even if I did not perfectly meet each goal (e.g. the discussion section was very poor in December), I nonetheless plodded steadily through this Capstone and had few moments of looming deadline frenzy.

#QUALITYDELIVERABLES:

This final Capstone submission has been edited twice (once for the December submission, and once for the February submission) following feedback from Prof Wilkins, proofread twice by me before this submission, and proofread as well by Ruby Lenard (we proofread each other's Capstones). In the Fall, my goal was to get all of my ideas on paper, and I was successful in doing so. However, rarely is a first draft well-composed, and it was therefore important to carefully edit each section and paragraph. My approach here was to reverse outline each section **and** paragraph, and then rewrite using the reverse outline. I created figures with `matplotlib` with `plt.figure()` and saved them directly to the desktop using `plt.savefig()`, ensuring that the figures in this document are of high resolution. This entire Capstone was written in LaTeX, allowing for hyperlinked references, footnotes, figures and tables for easy navigation and clear presentation of mathematical formulae.

#CURATION:

If this Capstone were a journal article aimed for publication, it would be much shorter. The literature review section would be particularly compressed. Though the primary goal of *my* Capstone is to investigate a self-directed research question, the goal of a Capstone *in general* is to showcase and synthesise my understanding of Minerva's HCs and LOs. Thus I aimed to strike a balance of the concision required of a scientific investigation, while also expanding the exposition where necessary to demonstrate my in-depth understanding of the topic and HCs. This is why my literature review is so extensive: it offers a comprehensive overview of the topic appropriate to the scope of a Capstone, but too long for a scientific article. There were, sometimes, certain aspects of an HC necessary for a strong application, but inappropriate to the purpose of its application in the body. For instance, justifying why I decomposed the various levels that interact to generate publication bias is important to a strong application of #LEVELSOFANALYSIS, but not necessary for demonstrating that there are levels that interact and that this renders the problem difficult to resolve through intervention. In moments like this, I reserved the HC-specific exposition for the appendix. Similarly, a detailed description of how the NUTS sampler works is irrelevant to the study's Methods (it's only relevant to state that I used a NUTS sampler), yet is important to #ALGORITHMS. Where an HC/LO has been applied in a specific place, I use footnotes to describe the specific application. If an HC/LO is applied in snippets throughout the paper (e.g. #NS125-QALMRI) I similarly use footnotes to signpost their application.

By keeping HC and LO-specific analysis in the appendix, I strive to ensure that the paper itself presents information relevant only to the study: the literature review informs the research question's relevance, and the rest of the paper pertains only to those elements important to the methods, results and inferences. The appendix allows me to demonstrate my application of HCs and LOs where an exposition in the paper itself would be inappropriate.

#OUTCOMEANALYSIS:

To ensure quality HC/LO applications, I relied heavily on the HC handbook. For each HC, I went through the list of guiding questions and common pitfalls as I wrote the relevant appendix entry. I then got a second pair of eyes on the application by feeding ChatGPT 4.0 with the relevant list of guiding questions and pitfalls, the paragraph(s) in which I applied the HC, and asked it to evaluate the paragraphs against them. See an example conversation here.[21] This was especially useful before I gained feedback from Prof Wilkins on my HCs for the December and February deadlines. Per Prof Wilkins' advice early in the Fall, I planned my HC/LO applications **before** embarking on the Capstone; I knew what HCs I would apply in what sections so that the applications were pertinent, rather than shoehorning them in after the fact.

#PROBABILITY:

I first used this HC in the sections on frequentist and Bayesian hypothesis testing (Footnote 2). My discussion of Fisherian and Neyman-Pearson hypothesis testing establishes the benefit of Bayesian methods over them. Specifically, I use this HC to describe the differences between Fisherian and Neyman-Pearson p values and describe why neither provides direct measures of evidence for an alternative hypothesis H_A. I then describe why Neyman-Pearson α values are incompatible with p values as they are currently used, and how these lead to grievous and widespread misinterpretation that opens a large doorway for bias and poor statistical technique to enter research. This required a nuanced consideration of what probability a p value represents, namely the probability of the data under H_0, and what α represents, namely the probability of a false positive.

Then, in my section on Bayesian methods, I describe how the Bayes' factor is computed mathematically and compare the Bayesian approach to the frequentist approach (Footnote 3). Given the goal of evaluating evidence in favour of some hypothesis H_A, I describe why the Bayesian approach provides a direct measure of this and improves on the flaws of NHT. My brief

[21] Feel free to share this application with other faculty, I did see your comment about this on the revised draft.

description of the analytical intractability in computing Bayes' factors in many cases reflects an understanding of continuous probability distributions and the challenges of integrating out over a multi-dimensional $\vec{\theta}$.

In my data analysis and discussion thereof, I draw an analogy from the rate of $B < 1$ to a frequentist Type II error rate. Since the studies were simulated with an implicit true effect, those Bayes' factors less than one would lead the researcher to falsely favour H_0. In the Discussion, I describe how the probability of a Type II error is greater with the Bayesian framework than with a frequentist hypothesis test and contrast this with the demonstrated robustness of B to prior hacking. In short, I identified, computed and interpreted the Type II error risk and discussed its implications for the overall research question.

#BIASIDENTIFICATION

This HC is critical in establishing why publication bias and p-hacking are both ubiquitous problems and are also here to stay. Clearly and robustly establishing this justifies the Bayesian approach's relevance.

In my discussion of p-hacking, I describe how confirmation bias, the sunk-cost fallacy, and vested-interest bias on the researcher level all contribute to a tendency to push positive results for publication and discard negative results into the file drawer (Footnote 7. I then describe how these biases permeate into a meta-analysis, and discuss how bias pervades even in those meta-analytical protocols intended to mitigate bias (Footnote 9

My analysis of publication bias and its sources also leans heavily on this HC (see #LEVELSOFANALYSIS): each level is subject to its own biases. I describe the possible biases that would influence editors to favour positive results via confirmation bias and the newsworthiness of positive results, why funding agencies, especially in the private sector, have financial incentives to prefer positive results, and why these factors (or a belief in them) leads researchers to both p-hack and to tend only to pursue their positive findings. It is the interaction of these biases across levels that makes these biases difficult to fight systematically and calls for a need for retrospective statistical techniques to identify and correct those biases.

#LEVELSOFANALYSIS

I leverage this HC in my discussion of publication bias and *p*-hacking to examine the various causes and factors within and across levels in the publication process that make it ubiquitous across research domains (Dickersin, 2005) (Footnote 6). I decompose the process into three tiers: funding agencies, academic journals, and researchers. I explain what biases lead each level to favour positive results and describe how these biases interact across levels (for instance biased funding sources pushing researchers to obtain positive results to ensure their funding). I explain how the pervasion of these biases across levels makes publication bias an incredibly difficult issue to solve (e.g. with a prospective meta-analysis), and this is an important motivator both for meta-analyses and for quantifying and mitigating those same biases within them.

Elemental to this HC is a nuanced discussion of interactions between levels.[22] For this discussion, a researcher has two primary goals: funding, and publication. To secure the former, the researcher interacts with the funding agency. To secure the latter, the researcher interacts with a journal, the key agent being the editor. Whether editors indeed favour positive results or not (Olson et al., 2002; Seward, 1999; Sutton & Pigott, 2005), they are the ones who decide whether a researcher's work is published or not, and it is, therefore, a researcher's *belief* about an editor's bias that matters. Indeed, researchers do tend to believe that journals (i.e. editors) will likely reject their negative results (Johnson & Dickersin, 2007; Joober et al., 2012). While newsworthiness and status are cited as reasons that editors may reject negative results, these factors are also likely to bias our researchers against submitting a negative result, since they believe these factors impact the editor, whether that is the case or not.

The funding agency's bias affects the researcher directly. A researcher, eager to secure funding, is more likely to secure funding when their existing research is favourable to the agency's goals (Eisenberg, 1988; Johnson & Dickersin, 2007). This is especially true when a researcher interacts with private funding agencies (Aghion et al., 2008; Eisenberg, 1988; Olson et al., 2002).

[22] Since the literature review is already extensive and not centred on bias sources, I've saved this discussion for the appendix.

Thus bias on the funding level pervades via interaction with researchers, while researcher biases *about editors* are problematic since publication is reliant on the researcher-journal interaction. Thus, one approach to mitigating publication bias would be to reduce the researcher-funder interaction (e.g. by expanding government funding or increasing regulation of private funding agencies), and working on the researcher level to combat the belief that editors favour positive publications. By targeting the interaction, rather than the bias at each level, we more effectively mitigate the bias. Since this is not occurring, though, we need, in the meantime, statistical techniques.

#BIASMITIGATION

The sections "meta-analysis" and "quantifying and detecting bias" are both focused on mitigating publication and bias.

I firstly discuss meta-analyses as a bias mitigation technique (Footnote 8. I describe how, by pooling studies together, meta-analyses provide much larger effect sample sizes, and how pooling intends to neutralise bias. I then explain how the one-sidedness of publication bias and *p*-hacking makes this a faulty assumption, thereby justifying that bias remains an issue in meta-analysis and support this with numerous studies on the issue (Ioannidis, 2005; Simmons et al., 2011; Wicherts et al., 2016).

After explaining why bias may still permeate a meta-analysis, I describe different approaches for mitigating that bias. I begin by explaining how prospective meta-analyses (Berlin & Ghersi, 2005) provide a framework with fewer gaps in which bias may permeate by formalising each component study specifically for meta-analysis, and then describe why this is an unfeasible approach (Footnote 11.

I then describe the primary statistical techniques used to quantify and mitigate that bias, as outlined in (Sutton & Pigott, 2005) (Footnote 12). All of the techniques I discuss are graphical or statistical techniques; they are not cognitive bias mitigation strategies. Thus I also justify why there is a need for such statistical methods, in that, save performing a prospective meta-analysis, there is no way to guarantee that every component study is free from bias; that every study on this topic, especially those with null results, was identified. I describe each technique and then discuss its limitations, which leads me back to Bayesian approaches and their advantages both as hypothesis tests and being bias-resilient.

#NS125-LITREVIEW

To motivate my question and justify both its need and theoretical relevance, I conducted an extensive literature review following the guidelines in Creswell (2014). The first step is to identify a topic; a short title to direct the literature review and identify a tractable and fruitful subtopic. This was the goal of my Capstone Proposal in CP192, where I first identified the topic "Bayesian modelling of *Z* curves." As I skimmed some of the literature in May, I settled on the broader "bias in meta-analysis."

From here, the next step was to evaluate if I could feasibly research the topic. After reading the simulation study from Simmons et al. (2011), I realised that I could conduct an original investigation right from my laptop.

Creswell writes that one should be able to justify whether their topic is worth studying before embarking on the literature review, but this also assumes enough knowledge of the topic to make that assessment. Knowing very little, my goal of the review was to identify and justify such a fruitful direction (see #NS125-FRUITFULDIRECTIONS).

I chose, then, to take an integrative approach to my literature review, with the goal of summarising the broad themes in the literature. This involved breaking down my topic (see #BREAKITDOWN) and using Google Scholar to identify sources. I started with an initial database of 50 sources which I skimmed (per Creswell's advice) to familiarise myself with the literature space. I then began my closer reading with systematic reviews to familiarise myself with the topics (Dienes (2021), Egger and Davey Smith (2009), Rothstein et al. (2005), and Wagenmakers and Grünwald (2006)) and identified further sources within their bibliographies. I also used tools like Connected Papers and ChatGPT (see the AI integrity statement, Appendix A) to survey the literature beyond my initial search and ended up with a database of around 160 sources. I kept a running research journal where I jotted down key points and ideas and referred to this continuously as I built the review.

After doing my reading, the next step was to construct a salient review (see

#ORGANISATION). Though my study is quantitative, Creswell's recommendation of organising a review by variable didn't fit with my goal of extracting a question *from* the literature: when I began, I had no specific variables to investigate. So, instead, I did another breakdown (see #BREAKITDOWN) that aimed to provide a narrative flow that built up to the question I had in mind (see #HYPOTHESISDEVELOPMENT, #NS125-FRUITFULDIRECTIONS, and #NS125-QALMRI).

Following the extensive review, I aimed to summarise how the key takeaways from the review of each topic together built a picture that a) Bayesian methods improve on traditional null-hypothesis testing methods and b) it is unclear how researcher degrees of freedom impact the accuracy of estimates in meta-analyses of these Bayesian hypothesis tests.

#EVIDENCEBASED

Footnote 15 describes how the evidence in the literature review builds up to my research question (see too #NS125-LITREVIEW and #ORGANISATION). Throughout the paper, I use multiple scholarly sources of various types to support my claims, often citing multiple sources for a single claim. For instance, I use mathematical sources to provide rigorous theoretical evidence to my claims (e.g. the theoretical flaws to NHT as described by Hubbard et al. (2003) and Schervish (1996)), alongside qualitative discussion of those same flaws from reputable scholars in the field (e.g. Berkson (1943) and Ioannidis (2005). Similarly, my arguments about the vulnerability of frequentist tests to publication bias are supported both with simulation studies demonstrating this (e.g. Simmons et al. (2011)) and qualitative, sociological investigations into the issue (Fanelli (2011) and Joober et al. (2012))

Supporting claims with scholarly evidence is itself insufficient to form a robust case. By providing a breadth of *types* of scholarly evidence (theoretical, empirical, simulations, psychological, responses to journal articles etc.), I craft a robust case backed by a variety of methodologies; if one study were disproved, this would weaken my case, but not disprove it.

#BREAKITDOWN

Before reading

The first step to researching the topic was to break it down into different sub-areas to read about. Critically, I had little idea where the project would end up, nor did I have a good idea of a specific sub-topic to focus on. My broad topic was "bias in meta-analysis." I first followed Creswell's recommendation, breaking the literature down by type:

1. Review
2. Journal article
3. Book
4. Conference papers
5. Dissertations

I started by using "bias in meta-analysis" as a search term in Google Scholar to identify some useful review articles which gave me a better idea of the literature landscape. After reading reviews from Dienes (2021), Egger and Davey Smith (2009), Rothstein et al. (2005), and Wagenmakers and Grünwald (2006), I broke down by theme, and returned to my first breakdown to read within each theme:

1. Hypothesis testing
   (a) NHT
   (b) Bayesian
2. Publication bias
   (a) In individual studies
   (b) In meta-analysis
3. P-hacking

   (a) Causes
   (b) Consequences

By first breaking down by type of literature, I was able to efficiently familiarise myself with my topic of interest and begin to understand the various sub-areas and directions that my project could take. Broadly familiarised with the topic, I felt equipped to do a thematic breakdown. By falling back on my first breakdown, I could delve deep into interesting and relevant areas by reading journals and book chapters, while starting with reviews to determine where a deeper dive would be necessary.

After reading

Though my breakdown above was effective for gaining a rich understanding of the literature, it does not necessarily translate to an appropriate organisational breakdown and structure for the literature review. With the goal of a narrative structure that eventually lead to a fruitful question grounded in the literature, I chose to break it down by how "applied" the question was. In this way, I would begin with the theoretical foundations of the topic and then walk the reader from there towards a much more specific question. The breakdown is as follows (and reflected in the headings of the review):

1. Hypothesis testing (theoretical foundations)
   (a) NHT
   (b) Bayesian
2. Bias in single studies
   (a) Publication bias
   (b) *p*-hacking
3. Bias when aggregating studies
4. Mitigating the bias

I begin with hypothesis testing to introduce the theoretical foundation of my topic. Identifying the flaws on this level motivates the relevance of the Bayesian approach (critical to the preamble to the question, see #NS125-FRUITFULDIRECTIONS) and is important to the later discussion of *p*-hacking. Since my focus is on how bias permeates into meta-analysis, it makes sense to first discuss how it permeates a single study before describing its aggregate impact. This breakdown informed the organisational structure of the literature review, see #ORGANISATION for a discussion of that side of the breakdown.

Modularisation

Repeatable, adaptable and expandable code ought to be modular, and I designed my simulation with this in mind. I divided the code into three stages: data simulation, Bayesian hypothesis test, and analysis of Bayesian hypothesis tests. Within each stage, I created functions for each discrete step of the process. For instance, rather than having one function that generated the data from a single study within a loop and then plotted all the results, I created three functions: one for generating the data for a single study, one for looping over it, and another for plotting the simulated data. This not only helps with debugging, as each step could be individually run and tested but also allows for the simulated data to be easily regenerated with different parameters by changing only the input to a single function call.

#NS125-WORKINCONTEXT

Important to justifying the relevance of a research question is identifying the practical implications both of not having an answer, and answering it. I provide some worrying examples where publication bias has had negative consequences in society (e.g. approving drugs for things they do not effectively treat (Roth et al., 2004)) and numerous examples of biased funding to demonstrate the scope of the issue. Then, I return to these issues in the preamble to my research question to justify its relevance. See specific applications in Footnotes 5 and 10.

#NS125-QALMRI

Question

My research question, motivated by the literature review, is whether meta-analyses of studies that compute Bayes' factors are indeed more resilient to statistical hacking (Footnote 17).

Alternatives

The question lends itself to two simple alternatives. Either Bayes' factors are more resilient, or they are not. This is an over-simplification, and what is more important to identify is *how* resilient it is, and to what extent bias propagates into a Bayesian meta-analysis. These alternatives are testable by the logic of the study and deliver discrete predictions as to how the meta-analytically estimated Bayes' factor will differ from the true one. The former is computed with a Bayesian model, from which credible intervals can be computed, and the latter is analytically derived.

Logic

The logic of the study should ensure that the results clearly delineate between the two alternatives. The independent variable is bias in the component studies, and I capture this with a gradient of biased priors, from the most biased crystal ball prior to the unbiased uniform prior. The dependent variable is the divergence of the meta-analytic Bayes factor from the true Bayes factor. Since the true Bayes' factor is known and the meta-analytical Bayes' factor is estimated via the same model, these are both calculable quantities. The predictions of the study (see #TESTABILITY) delineate between the Bayes' factor being resilient to bias or not, and, since the divergence may be quantified, also deal with evaluating the extent of this and allowing for comparison with NHT's sensitivity to biased component studies.

Methods

Application at Footnote 18. My methods section concisely describes how the studies were simulated, how the Bayesian models defined by the logic in the previous section were implemented and sampled, and how the marginal likelihoods and Bayes' factors were computed.

It is essential that the methods directly implement the logic of the study (so that the alternatives can then be assessed), and they do. Simulating data by drawing it from normal distributions, with a true effect size built-in, mimics the process of comparing two groups where there is a true difference in means between those groups. The Bayesian models under H_0 and H_1 reflect the null and alternative hypotheses accurately (see too entries for #MODELING and #DISTRIBUTIONS). Conjugate priors were chosen to minimise sampling error from the MCMC algorithm, which is especially important considering how many times a posterior was sampled (500 sampling rounds for the subset of the overall study presented in this draft). A flaw in the methodology, which I acknowledge and will improve on in the spring, was the use of a harmonic mean estimator to estimate the marginal likelihood. The unreliability of this approach jeopardises the results and ought to be replaced with a more reliable approach.

Results

Application at Footnote 19. I present my results concisely and support the prose results with figures. The results again pertain to the logic of the study, in that the logic relies entirely on the Bayes' factors calculated under each σ_{X_T}. I also present the results for the frequentist analysis as well as the meta-analytical p value for each $\Delta\mu$ group. The results and figures presented all directly pertain to the predictions described in the hypothesis (see too #TESTABILITY), and I interpret them in more depth in the discussion (see Interpretation)

Interpretation

Application at Footnote 20. I have interpreted the results concerning each hypothesis in the Discussion and used those interpretations to conclude that there is no significant advantage to the Bayesian methods, at least from the perspective of bias. I describe too the gaps in my investigation and suggest follow-ups and expansions that would address those gaps.

#NS125-FRUITFULDIRECTIONS

See Footnote 14. My research question is directly rooted in the literature review: I spend the review firstly motivating why NHT is theoretically suboptimal and prone to publication bias and *p*-hacking and then describe why meta-analysis is not a perfect solution to the problem. This motivates a need for Bayesian hypothesis testing, and I note the theoretical motivations for why it is less prone to bias (at least in terms of prior selection) but note a lack of empirical evidence to that end, despite subjectivity in prior beliefs being a frequent criticism of Bayesian methods.

My research question is therefore fruitful in that it builds off of an obvious gap in the literature, and its relevance is motivated by all of the flaws of NHT and meta-analysis that I spend the literature review describing. Given that Bayesian methods are posited as a solution to the crisis of false positives (Ioannidis, 2005), empirically investigating the truth of this claim seems quite important.

#HYPOTHESISDEVELOPMENT

See too Footnote 16. In my hypothesis, I establish a clear causal link using an "'If A then B because C" structure between the type of Bayesian prior and the level of Bayes' factor inflation. This relationship, supported by Dienes (2016), is directly amenable to my simulation approach since it only has me compute many B_is, to which both real and simulated data are equally conducive. By examining the impact of priors of diminishing widths the hypothesis provides measurable predictions and engages with the broader discussion on Bayesian test bias. The alternative hypothesis is that Bayes' factors are not resilient to bias, and this would be exhibited by a strong deviation from B in all groups that diminishes with growing prior width. The hypothesis lends itself to the logic detailed in the body and summarised in my description of #NS125-QALMRI. I also present a hypothesis that explicitly compares the Bayesian and frequentist approaches and offers two predictions that would support B's greater robustness to bias.

#TESTABILITY

I provide a testable hypothesis with multiple predictions that focus on the robustness of B to bias, and what both high and low degrees of B inflation say about this. I also provide the prediction that the inflation will be lesser with wider and less informative priors, which, in being less informative, are less biased, which provides another, more nuanced line of evidence. A separate line of evidence is in the proportion of B that rejects the alternative: a higher proportion suggests a more conservative test. These predictions are testable in practice (see Logic in #NS125-QALMRI. In the overarching context of my paper, these predictions also lend themselves to comparison with the inflated effect sizes that are observed in NHT meta-analyses (Friese & Frankenbach, 2020; Ioannidis, 2008; Simonsohn et al., 2014b).

We may also make predictions as to what disproves the hypothesis. These would be predictions that demonstrate that Bayesian methods are as or less robust than frequentist methods of statistical testing. Some such predictions are:

- $\bar{B}$ above zero for biased priors for small $\Delta\mu$ (Figure 7)
- Distribution of B not symmetrically distributed about $B = 0$ (Figure 6)
- Lower change in false negative rate with meta-analytic p than $\bar{B}$ as $\Delta\mu$ decreases.

#COMPARISONGROUPS

To capture the varying degrees of bias in prior selection I compare normal priors over the effect size $\Delta\mu$ of varying widths by adjusting σ. The widest $\sigma = 5$ prior is the least biased group, although it may be valuable to also incorporate a totally unbiased uniform prior, though its domain is not all of $\mathbb{R}$ as it is for the normal priors. None of these comparison groups represent a realistic study sample for a meta-analysis. In reality, a meta-analysis would incorporate studies that all used different types of priors, each with varying degrees of bias, and those studies would be biased in different ways (choice of likelihood, posterior sampling and reporting etc.) Nonetheless, capturing all of the nuance of this bias would be exceedingly complex, and unnecessary to the hypothesis. These comparison groups, varying in width, nonetheless capture varying degrees of bias, especially the very narrow $\sigma = 0.01$ prior, which is nearly a crystal ball prior. Demonstrating that even the $\sigma_{X_T} = 0.01$ is robust implies that Bayes' factors are robust too to less-severe hacking. These simple comparison groups directly test the hypothesis, even if they are not realistic.

#PLAUSIBILITY

The key assumption of the hypothesis is that despite priors being a major area of researcher freedom, given a large enough sample size (enough studies and enough participants), the Bayes' factor derived from the posterior is accurate regardless of the choice of prior. This is a fact of Bayes' equation: the more data there is the less important $P(H)$. This assumption is well-discussed in both Dienes (2016) and Wagenmakers and Grünwald (2006), and it is a reasonable one. Yet it remains a point of contention (thus the need for this study!) and I present an alternative hypothesis that would discredit it.

#MODELING

Tractably simulating B hacking required the use of models at all stages of the process.

Model one: Studies

To estimate the Bayes' factor for a set of studies, a set of studies are necessary. I model the simple case of a two-group difference of means test between a placebo and treatment group (i.e. the most basic kind of RCT) by sampling outcomes from two normal distributions, each with fixed means. This allowed me to encode a true effect size while also ensuring variation between studies. Importantly, since we are interested in a non-significant effect, I choose a small effect size such that the data appears insignificant under frequentist testing. The model assumes a normal distribution over outcomes, which mightn't be true for all types of studies. It also assumes that the effect size is the same across studies. This is an unrealistic assumption, as though studies *aim* to investigate the same effect, it is nay impossible to perfectly replicate a study, and the true effect size under each study design will vary. One way to model this would be to distribute the effect size too when generating the data. Nonetheless to the end of comparing Bayesian models for a simple difference of means test, this model effectively generates data with differing means while still allowing for stochasticity.

Model two: Hypotheses

The second model was that of the hypotheses. This is really two models, one for H_0 and one for H_1. Here are the models again. For H_0

$$\begin{aligned}
\mu &\sim \mathscr{U}(-1,1) \\
\Delta\mu &= 0 \\
X_O &\sim \mathscr{N}(\mu,1) \\
X_T &\sim \mathscr{N}(\mu+\Delta\mu)
\end{aligned}$$

And for H_1:

$$\begin{aligned}
\mu &\sim \mathscr{U}(-1,1) \\
\Delta\mu &= P_n(\Delta\mu, H_1) \\
X_O &\sim \mathscr{N}(\mu,1) \\
X_T &\sim \mathscr{N}(\mu+\Delta\mu)
\end{aligned}$$

These are Bayesian models that attempt to encapsulate two hypotheses over the data-generating process: either there is no true effect size, or there is one. I use a flat and wide prior over μ to be as unbiased as possible, though μ is known and defined from how the studies were simulated. It was also important for this model to contain an RDF wherein bias could penetrate, and this is in the prior over $\Delta\mu$. Another key element of the model is that it minimises the risk of biased sampling since PyMC uses MCMC sampling. Using normal distributions as both the likelihoods and priors ensures a normal posterior (since the conjugate to a normal is normal). The model's priors place some implicit assumptions. The first is that $|\mu| < 1$. Since the simulated data satisfy this assumption, this isn't an issue, but adapting this to a real set of studies would require a more informed prior over μ. The second is that X_O and X_T are independent. Though RCTs are the gold standard in ensuring this, not all studies follow this standard rigorously.

Model validity

It is essential that the model accurately captures the key elements of a real Bayesian hypothesis test while making the appropriate assumptions to be efficiently simulated.

While there do exist guidelines for Bayesian versions of common frequentist hypothesis tests (e.g. t-test) (Wagenmakers & Grünwald, 2006), there is no unanimously agreed upon standard as there is in frequentist statistics. My model presents a very simple conceptualisation of a difference of means test where the mean and standard deviations of each group are unknown. The kernel here is to show through simulation that the choice of prior has a negligible impact on the overall Bayes' factor, and this is the simplest model of a difference of means test. Though more complex hierarchical solutions do exist (e.g. Fong (1989)), simplicity is generally better, and I would rather not introduce more RDFs into my work by opting for a hierarchical model.

In terms of the priors used, these are standard choices for this type of problem. A normal prior over $\Delta\mu$ is sensical given that the CLT guarantees that $P(\Delta\mu) = \mathcal{N}(\mu, \sigma^2)$ in the limit of infinitely many samples. Narrowing this prior is the simplest way to simulate varying degrees of researcher certainty (or bias) in the true value of $\Delta\mu$. Moreover, the prior may also be chosen given information from previous studies, in which case it would inform the choice of μ and σ that goes into $P_n(\Delta\mu)$.

Model three: Researcher bias

The key process I model in this paper is researchers performing biased Bayesian statistical tests. There are many researcher degrees of freedom in a Bayesian statistical test, more than in a frequentist one. In addition to bias in the data collection and processing phases, a researcher builds the Bayesian model themself, choosing the model parametrisation, prior(s) et cetera. Modelling all of the different combinations of bias amongst these degrees of freedom would be far beyond the scope of a Capstone. I focus on biased priors, a) because the freedom of prior selection is a common criticism of Bayesian hypothesis tests (e.g. (Begg & Berlin, 1988; Dienes, 2016)) and b) the prior represents a researcher's belief about their hypothesis and therefore is a place in which to encode bias. Thus this study makes the flawed assumption of perfectly unbiased data collection processes. As I discuss in the conclusion, this is a salient area of further investigation.

#DISTRIBUTIONS

This HC was applied in tandem with #MODELLING. I model treatment outcomes X_O and X_T with normal distributions. As the maximum entropy distribution assumes fixed mean and variance, this is an appropriate choice for modelling the general case of a treatment variable we know nothing more about. I leverage the fact that the conjugate prior to a normal distribution is also normal in order to minimise the risk of sampling error with MCMC. I discuss the assumptions imposed by choosing these distributions in my discussion of #MODELING above.

I also apply this HC in building my simulations of the data. I model each study as sampling the change in a treatment outcome. Whether the dependent variable is continuous or discrete is unimportant,[23] since we are measuring a difference over averaged quantities. Thus it is appropriate to model it as $\mathcal{N}(0,1)$ (again as it is the maximum entropy distribution for fixed μ and σ).

#CS146-PYTHONIMPLEMENTATION

I use modularised code to simulate the data, compute the Bayes' factors over it under each model and prior variation, and plot the results; rather than implementing `PyMC` models directly in a code cell, I defined them within functions. When I expanded my study in the spring, this allowed me to easily re-run the models over different datasets, without needing to re-define models over and over again. Implementing the models in `PyMC`, as we were taught to do in the course, reduces the margin for mathematical error by handing over the mathematical implementation and sampling of the models to the package. I describe in detail how the package samples and evaluates Bayesian models in the entry #ALGORITHMS.

[23] Provided that the discrete quantity is ordinal and that decimal values remain sensical

#ALGORITHMS

I produce efficient and adaptable code that is divided into discrete functions. This is far better than writing straight code into a code cell since it means that I can easily change the simulated data by altering the parameters to the function that generates the simulated data, and similarly can study the effects of different priors, effects etc. by changing single lines of a function, rather than needing to search the code and update a variable every time it is called. Functions also aid in readability and ease of understanding of the code since they naturally compartmentalise the program into its different functions.

To implement, sample and extract data from the Bayesian models, I used the `PyMC` package with a NUTS sampler. The package firstly mathematically assembles Bayesian models into a `Model` object. We sample from the model's various distributions with `pm.sample()`.

We sample with a Hamiltonian No U-Turn Sampling algorithm. The algorithm explores the posterior (or likelihood, posterior-predictive, prior etc.) by modelling a particle moving frictionlessly upon the (flipped) posterior distribution. We conceptualise the particle as having some constant energy $E(\theta,m)$ (since the surface is frictionless):

$$E(\theta,m) = \mu(\theta) + K(M)$$

Where μ is the particle's gravitational potential energy defined by its location on the negative log posterior (dependent therefore only on the parameters θ), and K the kinetic energy dependent on the particle's momentum m:

$$k(M) = \sum_{i=0}^{N} \frac{M_i^2}{2m}$$

$$\mu(\theta) = -\log(P(x|\theta)P(\theta))$$

k is computed as the sum of the energy in each dimension of posterior space, while the μ is defined by the negative log of the unnormalised posterior (as the posterior probability density increases, the GPE gets lower and the particle 'falls' further).

The probability distribution over the particle's energy is defined by the Boltzmann distribution:

$$\begin{aligned} P(E) &\propto e^{-\frac{E(\theta,M)}{T}} \\ &= P(x|\theta)P(\theta)e^{-\frac{M^2}{2}} \\ &= P(x|\theta)P(\theta)\mathscr{N}(M|0,1) \end{aligned} \tag{5}$$

The last two steps of the above equalities assume the particle's mass is 1, and that the system's temperature is 1K. By sampling from the distribution defined in (5) and discarding momentum samples (M), we end up sampling from the posterior.[24] To generate these samples, we sample from the normal momentum distribution and use Newtonian mechanics (and the equations for k and μ) to solve for the particle's path given a random start point on the posterior and our randomly drawn initial momentum. The sampling algorithm then proceeds in the conventional Metropolis-Hastings manner:[25] along a sampling chain (I used 4 chains each with 2,000 samples), a new point $P(\theta^*, M^*)$ is proposed, and accepted or rejected by computing:

$$r = \frac{P(x|\theta^*)P(\theta^*)\mathscr{N}(M^*|0,1)}{P(x|\theta)P(\theta)\mathscr{N}(M|0,1)}$$

The proposed point is accepted if $r > u$ with $u \sim \mathscr{U}(0,1)$.

Hamiltonian Monte-Carlo sampling is a strong sampling algorithm for posterior distributions, especially high-dimensional ones. The major advantage of Hamiltonian Monte Carlo is that it offers reduced autocorrelation compared to other MCMC methods; fewer samples are needed to sufficiently explore the posterior.

The NUTS sampler incorporates a no U-turn condition on the HMC sampler. For a given sample, we continue to take discrete trajectories until the next proposal backtracks over us. This

[24] This is easily demonstrated by marginalising out M.

[25] The Hastings term is dropped, since we assume Hamiltonian movement to be deterministic between the previous and proposed points in both directions

further reduces computing time, as the number of steps necessary to arrive at a given final point is not fixed; we better explore the posterior by not performing u-turns as we sample a given trajectory.

NUTS' disadvantage is that it is slow. Sampling of the posterior for a single hypothesis for a single study takes about 23 seconds, and this is done for two hypotheses, the latter with five different prior parametrisations for 50 studies, for five different $\Delta\mu$s. This part of the code could be optimised, not by changing the algorithm, but rather by tuning it. Nonetheless, for my purposes, I had the time to run the code cells overnight and prefer to be sure of my output, rather than tuning it and risking inaccurate results.

#SIGNIFICANCE

The primary result from my investigation is that Bayesian hypothesis tests present a more reliable estimate of a hypothesis' truth than a frequentist test of significance. I devote the first part of my literature review to describing firstly how Fisherian and Neyman-Pearson significance tests examine a hypothesis' truth from a theoretical standpoint and then unpack why the modern approach to significance testing is flawed both practically and theoretically. I then describe precisely where frequentist tests are prone to bias and manipulation, relating significance values to Type I error rates. By approaching significance testing from its foundation (what is a *p* value, how does it relate to Type I error), rather than going into detail on specific significance tests, I lay the stage to contrast the common frequentist approach to the Bayesian one. Pitting frequentist and Bayesian methods together like this lays the groundwork for my research question and the subsequent investigation.

#ORGANISATION

The organisation of the literature review was crucial in justifying the research question's importance. I aimed to create a narrative arc that began with setting the scene: what is hypothesis testing, and what tools are available (Fisherian, Neyman-Pearson, and Bayesian). I then establish why Bayesian methods are posited as stronger by first describing the theoretical flaws of frequentist methods, and then explaining how Bayesian methods address those flaws. Having established the theoretical improvement offered by Bayesian approaches, I then introduce issues of publication bias and *p*-hacking (a practical, rather than theoretical, problem), and again describe why Bayesian methods are thought more robust to these errors. This set the stage for my research question. By dividing my literature review into distinct but interconnected topics, and structuring it such that the entire review builds up to the research question, I motivate both the subject-specific and real-world relevance of my study (see also #NS125-WORKINCONTEXT and #NS125-FRUITFULDIRECTIONS).

As for the organisation of the overall paper, I organise it like a typical scientific article (though the literature review is much longer than would be typical). I place the literature review first again to provide the necessary background for, and to motivate, the topic of investigation. I present the methods concisely and leverage my appendices to show the code and HC/LO relevant exposition.

Each paragraph and section was also written with organisation in mind. See Footnote 1 and 4 for exemplary discussions of paragraph and section organisation respectively.

#PROFESSIONALISM

I have formatted the entire paper in APA 7 style, as I would need to do so to publish in *Psychological Science*. I have written in an academic register and strived to use the active voice. The entire paper is formatted in LaTeX, allowing not only for clear typesetting of mathematical symbols and equations but also for each citation, figure and equation to be hyperreferenced, allowing the reader to quickly jump to the relevant bibliography entry. Moreover, bibliography formatting is handled by BibTeX, so there is a smaller chance that my bibliography entries are incorrectly formatted (though this ultimately depends on the metadata in Zotero)

I have chosen to write in British English a) as it is the register I use naturally and am therefore most likely to use consistently and b) out of personal preference. While my paper is formatted in APA style, the APA does not specify that American English must be used, and APA is a common convention used in psychology in the UK too. While using a style that originated in the UK (like Oxford style) might be more true to writing in British English, such a convention is not common in statistics or psychology. Thus I've stuck to APA.