Bayesian Hypothesis Testing: Sequentially Analyzing Real-World Data

Ardern-Mulhern, Thomas

Universiteit Leiden

Faculteit der Sociale Wetenschappen

Bayesian Hypothesis Testing: Sequentially Analyzing Real-World Data

Master's Thesis

Thomas Ardern-Mulhern

Master's Thesis Methodology and Statistics Master

Methodology and Statistics Unit, Institute of Psychology,

Faculty of Social and Behavioral Sciences, Leiden University

Date: July 2022

Supervisor: Dr Tom Heyman

Acknowledgements

I'd like to thank my Thesis supervisor Dr Tom Heyman for his support and patience as I worked my way through this project. For my family, who have remained supportive and encouraging from across the English Channel, without them I couldn't have even considered a Master's degree, let alone one in a foreign country. I'd also like to thank my partner Jildou, whose strength of will bolstered my own and helped me escape from orbit.

Table of Contents

Abstract

Prior research has compared Bayes factors and *p* values within Hypothesis testing using *t* tests (Wetzels et al., 2011). The current research expanded on this comparison to include both *t* tests and various forms of Analyses of Variance. Further, we conducted maximal *n* sequential analyses, following the design as proposed by Schönbrodt and Wagenmakers (2018). We conducted two forms of sequential analysis: The fixed order sequential analysis, in which the order the data was presented in the dataset dictated the order by which it was added, and the replicated random order sequential analysis in which the order was randomized, and the procedure repeated 100 times. For both the comparison and Sequential analyses we used Bayesian alternatives to Classical Significance tests with real-world data that were reproduced with author's assistance by Hardwicke et al. (2018). We found that Bayes factors and *p* values covary in both *t* tests and Analyses of Variance. However, we observed Bayes factors that underemphasized the perceived effect by *p* values, as well as Bayes factors that overemphasized when compared to *p* values even after performing a sensitivity analysis. We also found that most Sequential analyses produced Bayes factors exceeding a threshold prior to the maximal *n*, with most analyses exceeding more. We also contextualized our fixed order sequential analysis using the percentage of Bayes factors across the 100 replications of each sequential analysis that exceeded thresholds. We evaluate these findings and propose measures researcher may take based on our findings to utilize optional stopping in a way that is efficient, reasonable and accurate.

Bayesian Hypothesis Testing: Sequentially Analyzing Real-World Data

Conclusions within Psychology are supported via statistical inferences based on data. The most common form of statistical inference in psychological research are statistical tests that produce a tes*t* statistic such as an *F* statistic or a *t* value, which in turn is used to calculate a *p* value that represents the probability of encountering a test statistic at least as extreme as the one calculated given the null hypothesis is true. Devised by Fischer (1890–1962), this is known as null hypothesis significance testing (abbreviated to NHST or 'null-hypothesis testing' within this research). Because this form of statistical inference relies on results being compared to an initial hypothesis that there is no effect of interest, a smaller *p* value implies a greater likelihood of the alternative hypothesis being true, with a *p* value below a normative significance level of .05 being considered significant (Greenland & Poole, 2013).

Null Hypothesis Significance Testing

As discussed by Rouder, Speckman, Sun, and Morey (2009) when referring to *p* values derived from *t* tests, one cannot infer that a *p* value greater than a significance level (usually .05) is evidence for the null hypothesis being true. This is due to the relationship between the *p* value and the sample size of the data it is produced from. When calculating a *t* test for example, if the null hypothesis is false, then as the sample size increases, the test statistic *t* value becomes larger and the *p* value converges to 0. Therefore, when the null hypothesis is false, a larger sample size reduces the likelihood that the null hypothesis will be incorrectly accepted. When the null hypothesis is true however, a researcher cannot comparatively increase their sample size to increase the likelihood that the null hypothesis will be correctly accepted. This is because there is no upper bound for how large a *t* value can be, leading to a normal distribution which when transformed leads to uniformly distributed *p* values – a format which does not make the likelihood of one *p* value more likely than another (Nickerson, 2000).

Null-hypothesis testing is therefore flawed when used in large-sample research. In this case, the consistency by which data can be accurately assessed is not identical for both the null and alternative hypotheses. When the null hypothesis is false, the resulting null-hypothesis tests with a large sample will consistently (and accurately) reject the null hypothesis. When the null hypothesis is true however, significance testing results in a less consistent conclusion (Rouder, Speckman, Sun, & Morey, 2009). It is conceivable to assume that null-hypothesis testing has and continues to produce biased evidence against the null hypothesis, particularly if the sample sizes are larger than 30 participants. Coupled with the greater weight Psychologists place on the conclusions of research with larger sample sizes (Marszalek et al., 2011) this suggests that we should consider alternative null-hypothesis testing methods.

Classical significance tests are derived from the underlying logic of probability as the true and accurate representation of the "long-run relative frequencies [of events occurring]" (Dienes, 2011, *p*. 275). This interpretation of probability views outcomes within research as the same probability of those outcomes occurring within an infinite number of repeated experiments, this is the Frequentist perspective. An alternative perspective is that probability should be viewed as a quantified degree of an individual's belief in a particular outcome or model being true. In this alternative perspective, probability is not an absolute, it is conditional, and should be viewed as an abstraction by which we model uncertainty, this is the Bayesian perspective.

Bayesian Hypothesis Testing

In a paper by Reverend Thomas Bayes published posthumously, he argued that "The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon it's happening" (Bayes., 1763, *p*. 376). Therefore, the probability of an event can be

considered as the ratio of the odds of an event occurring. Within the Bayesian framework this notion of 'odds ratios' are considered within three main components when computing probability: the Prior odds, the Posterior Odds and the Bayes factor.

The odds for the models or hypotheses prior to data collection are referred to as prior odds. Within this research this is calculated as the probability of the Null Hypothesis (H_0) over the Alternative Hypothesis (H_1) as shown below:

$$Prior = \frac{P(H_0)}{P(H_1)} \quad (1)$$

Our prior odds will be set to 1 in this research, which means that prior to data collection we view both H_0 and H_1 as equally likely. The degree of belief derived following the collection of data is referred to as a posterior belief, within this research this is calculated by:

$$Posterior = \frac{P(H_0)}{P(H_1)} \times \frac{P(D|H_0)}{P(D|H_1)} \quad (2)$$

As illustrated above, calculating Posterior probability required the probability of the data given the hypothesis being true, this is known as the likelihood. While classical significance testing allows one to calculate the likelihood of the hypothesis given the data, $P(H|D)$. Due to the inability of classical significance testing to calculate the probability of a hypothesis, as well as classical significance testing's incorporation of decision procedures that do not influence likelihood (Dienes, 2011), this cannot be inverted to calculate the likelihood of the data given the hypothesis, $P(D|H)$. With Bayesian statistics however, application of probability to hypotheses (or indeed, anything) is acceptable. From this, one can calculate a ratio representing the probability of the data given the hypothesis by using subjective prior odds. The resulting calculation of marginal likelihoods produces a factor by which we may accept or reject our null or alternative hypothesis given the data. This is known as the Bayes factor, which is calculated in this research as:

$$Bayes\ Factor = \frac{P(D|H_0)}{P(D|H_1)} \quad (3)$$

It is important to note that within Bayesian Statistics, the odds ratios are sometimes inverted, such that the *Prior* = $\frac{P(H_1)}{P(H_0)}$, *Posterior* = $\frac{P(H_1)}{P(H_0)} \times \frac{P(D|H_1)}{P(D|H_0)}$ and *Bayes factor* = $\frac{P(D|H_1)}{P(D|H_0)}$. This means that a Bayes factor of two indicates that H_1 is more likely than H_0 given the data by a factor of two. However, to allow for an easier visual comparison to *p* values, the Bayes factors within this research were inverted such that values closer to zero indicated greater support for H_1 and values greater than one indicated support for H_0, a Bayes factor of 1 still indicates no support for H_0 over H_1, or vice versa.

Priors are subjective, researchers however have argued that assigning the values of priors for parameters should not be arbitrary. As explained by Rouder et al (2009), Bayes factors are ratios expressing marginal likelihoods that are calculated by:

$$P(D|H) = \int_{\theta \in \Theta_H} f_H(\theta; y) p_H(\theta) d\theta \quad (4)$$

Rouder et al argued that this marginal likelihood should be viewed as a continuous average in which the priors (p_H) are weights. As a result, "the prior should not attribute undue mass to unreasonable parameter values" (Rouder et al., 2009, *p*. 228). For one sample *t* tests, setting the mean of H_0 (denoted within this paper as μ_0) to 0 was viewed as a reasonable metric. assigning a particular value to the mean of prior H_1 (denoted within this paper as μ_1, which is a departure from Rouder's notation) is potentially problematic, as μ_1 priors that increasingly deviate from the data produce a BF that increasingly favors H_0. Instead of assigning a particular value to the μ_1 therefore, Rouder proposes applying a distribution to the μ_1, as shown below:

$$\mu 1 \sim Normal(0, \sigma^2_{\mu_1}) \quad (5)$$

However, as with the μ_1, applying a singular value to the Standard Deviation of μ_1 (σ^2_{μ}) leads to the same pitfalls as with applying a particular value directly to the μ_1. An

alternative proposed by Rouder and others is to determine a prior for the effect size, denoted as δ and calculated by $\delta = \mu 1/\sigma$. They suggested the following:

$$\delta \sim Normal(0, \sigma_\delta^2) \quad (6)$$

The advantage of parameterization of δ over μ_1 is that researchers can avoid placing too much weight on very large effect sizes. Again however, the Standard Deviation of the effect size, σ_δ^2, must be specified. Researchers have recommended an inverse Chi-squared distribution (Zellner, & Siow, 1980), allowing for the σ_δ to be constrained to 1 (Gelman et al., 2013). This inverse chi-squared distribution is equivalent to a Cauchy distribution, when paired with a selection of the prior for the variability of the data (σ^2) as $p(\sigma^2) = 1/\sigma^2$. This is referred to as the Jeffreys prior (Jeffreys, 1963). When both the Cauchy distribution for the δ and the Jeffreys prior is applied to the σ^2 this is known as the Jeffreys-Zellner-Siow, or JZS, prior. A JZS prior is also useful, as the Cauchy distribution can be adjusted using a scale parameter *r*, this allows researchers to refine their δ, with a larger r being more suited to a larger expected δ as denoted below:

$$\delta \sim r \times Cauchy \quad (7)$$

While Rouder et al recommended a default *r* of 1. For *t* tests, this research instead used the default *r* from the *BayesFactor* package of $\frac{\sqrt{2}}{2}$ (Rouder et al., 2009) as well as a default *r* of 0.5 for ANOVAs (see Rouder, Morey, Speckman, & Province, 2012 for more details). The use of this default prior and its potential disadvantages are explored within the Discussion section.

For clearer interpretability, researchers have categorized Bayes factors relative to their evidential strength in favor of a hypothesis. For example, Kass and Raftery (1995) categorized Bayes factors exceeding 1 as anecdotal evidence, 3 and 1/3 as substantial evidence, 10 and 1/10 as strong evidence. This research will be using these thresholds proposed by Schönbrodt & Wagenmakers (2018; see Table 1). These thresholds were adapted

from Wetzels et al (2011), which also labelled *p* values by similar thresholds to facilitate the comparison with Bayes factors. This research will include the threshold of 6 and 1/6 to nuance Bayes factors that do not exceed 10 and 1/10. We will label this additional threshold as 'promising'.

Table 1

Thresholds as adapted from Schönbrodt & Wagenmakers (2018)

Evidence for H_0 (BF > …)	Strength of Evidence	Evidence for H_1 (BF < …)	Evidence against H_0 (*p* values)
100	Decisive	1/100 (.001)	<.001
30	Very Strong	1/30 (.033)	
10	Strong	1/10 (.010)	.001 - .01
6	Promising	1/6 (.167)	
3	Substantial	1/3 (.333)	
1	Anecdotal	1	.01 - .05

In 2011, Wetzels and others ran an empirical comparison of *p* values, effect sizes and Bayes factors using 855 *t* tests gathered from contemporary research in Psychology (Bayes factors were calculated using the prior specification put forth by Rouder et al., 2009, henceforth default Bayes factors). They found that effect sizes, Bayes factors and *p* values all generally covary. However, the largest degree of covariance appears to be between *p* values and the default Bayes factors produced by Bayesian hypothesis testing. Wetzels et al. described the primary difference between *p* values and default Bayes factors to be "one of calibration: *p* values accord more evidence against the null than do Bayes factors" (Wetzels et al., 2011, *p*. 295). The current research intends to further investigate the nature of this relationship by comparing *p* values and Bayes factors conducted using default Bayesian alternatives beyond *t* tests such as Analyses of Variance (ANOVA) and Analysis of Covariance (ANCOVA). Beyond these modifications, this research will also be using a procedure simulating optional stopping.

Optional Stopping and Sequential Testing

As previously stated, classical significance tests calculate the probability of the data given the hypothesis and decision procedure(s). These decision procedures apply predetermined conditions to data, such as α and β error rates, as well as predetermining conditions for data collection. Decision procedures which are not predetermined are often not useable within the NHST framework. One example is the practice of optional stopping, this is when data collection is paused and examined periodically and resumed or stopped depending on the researcher's desired outcome. This is because of the long-term perspective of classical significance testing. If one considers the first test conducted on data to have an α error rate of .05, meaning that 5% of the time a significant finding is found that is due to chance, further testing will increase this error rate. Increases to the α error rate would eventually indicate that a significant value is inevitable even if H_0 is true. Within an NHST framework this is considered bad practice and is seen as an example of '*p* hacking' (Simmons, Nelson, & Simonsohn, 2011). Alternative null-hypothesis testing using Bayes factors however does not consider more spontaneous decision procedures to be problematic (Dienes, 2011). Decision procedures such as optional stopping are therefore perfectly suitable within the Bayesian framework.

Optional stopping is useful for maximizing a researcher's efficiency during data collection. Using Bayes factor design analysis, Schönbrodt and Wagenmakers conducted Bayesian alternatives to power analyses, in which they compared three experimental designs in which Bayes factors were produced in order to determine which design was the most informationally rich and maximally efficient. These were a fixed-*n* design, where a sample size was pre-defined, and the analysis conducted once all the data was gathered – this did not include any sequential testing. Another was an open-ended sequential testing design, where Bayes factors were calculated as participants were added, until a desired level of evidence is

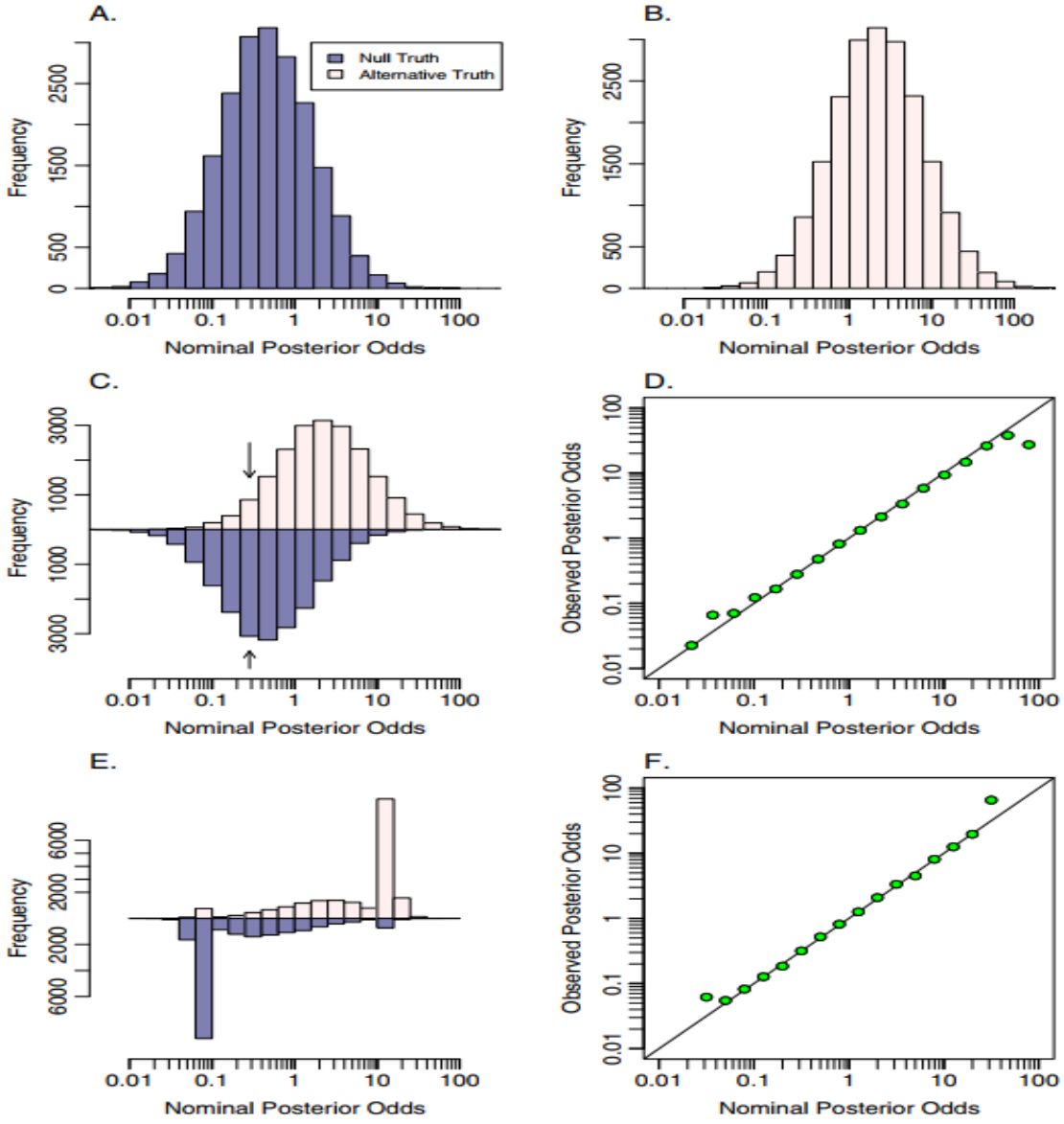
reached. Finally, they also considered a sequential design with a maximal *n*, where participants were added until either a sufficient level of evidence is reached, or the pre-defined maximum number of participants was added. They recommended that researchers use a sequential design with a maximal *n*, as it can allow researchers to stop data collection if sufficient thresholds are crossed while acknowledging the reality that researchers have limited resources (Schönbrodt & Wagenmakers, 2018).

Some researchers have suggested that optional stopping is a potential problem within Bayesian analysis, producing "substantial irregularities" and that Bayesian hypothesis testing if used to reanalyze previously collected data is susceptible to an optional-stopping-rule artifact (Yu, Sprenger, Thomas, & Dougherty, 2014, *p*. 280). Others have argued counter to this, pointing out that Bayes factors are impervious to a modified data-collection technique due to the differences in interpretation between Bayesian and Frequentists perspectives on probability (Rouder, 2014). Within the Frequentist perspective, a significant *p* value is an immutable declaration of significant findings, as significant results reflect the conditions of a larger population. Because of this immutability, worries as to a more subjective method of data-collection are naturally problematic. A Bayes factor however makes a far more subjective claim. The conditions by which the data is collected is therefore far less important. Rouder (2014) examined the interpretability of Bayes factors. By simulating 20,000 replications of 10 observations created under one of two hypotheses – one in which the effect size is 0 (a typical prior for H_0) and the other where the data created had an effect size of 0.4, this was H_1. These 10 observations were created under one of these hypotheses as though under a fair coin-toss, meaning that the prior odds for data generation were set at $\frac{1}{1}$. Once the data was gathered, Rouder examined the posterior odds that were produced (which were equal to the Bayes factor as the prior odds were 1). Rouder observed that within this simulation, the replicated experiments yielded observed posterior odds based on the ratio of

H_0 and H_1 observations that were generated which agreed closely with what was predicted. This effect even held when the conditions of the simulation were changed such that sampling continued until a pre-determined Bayes factor was reached or sampling reached its maximum. This produced a distribution of nominal and posterior odds that deviated from a normal shape, as shown below in Figure 1.

Figure 1

Distribution of Posterior and Nominal Odds from Rouder (2014).



Note. Adapted from *Optional stopping: No problem for Bayesians*, by J. Rouder, 2014, p. 304.

While some researchers concluded that this meant optional stopping was problematic (Sanborn & Hills, 2014), Rouder argued that that the posterior odds remained interpretable,

as the "critical question is whether the posterior odds accurately reflect the probability that a given value came from a given hypothesis" (Rouder, 2014, *p*. 305). Other researchers however have taken issue with this conclusion (De Heide & Grunwald, 2021), these contrasting views will be expanded on in the discussion section.

The conclusions of Schönbrodt and Wagenmakers' research and Rouder's however are potentially limited by their use of simulated data. It is unclear if such Bayesian design analyses were applied to real data, in what instances sample sizes below their maximal *n* would exceed informative thresholds, or indeed how often these Bayes factors would exceed thresholds matching that of the maximal *n* or even indicate support for the same hypothesis.

Our aims within this research aimed to investigate the accuracy of Bayesian hypothesis testing. We aimed to do this by expanding on research comparing *p* values and Bayes factors, by using both real-world data and expanding the analyses that were compared to include both *t* tests and Analyses of Variance. We also wished to assess Bayesian hypothesis testing within an optional stopping framework. This was done by employing Schönbrodt and Wagenmakers' (2018) maximal *n* sequential design. For this research we employed two types of maximal *n* sequential designs. For the first type of design, we mimicked the conditions the researchers themselves would have experienced during the data collection by conducting a sequential analysis in the way in which the researchers collected their data. We referred to this as the fixed order sequential analysis. We also conducted a type of sequential analysis to both contextualize our findings in the fixed order design, as well as assess whether our findings remained consistent if the data was collected in a different order, we referred to this as a replicated random order sequential design. Using these sequential analysis designs, we aimed to calculate the retroactive efficiency gain by simulating the impact of optional stopping on real world data using the fixed order sequential analysis and

evaluate the consistency of these findings using the replicated random order sequential design.

Method

Materials

This research was conducted using secondary, real-world data from Hardwicke et al.'s paper (2018) "Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition". Datasets for which the initial author's results were successfully reproduced with assistance from the original author were selected for this study. Of these 11 datasets, two were excluded from this research, one because its analysis already included a Bayesian test and the other used statistical testing which specified a Greenhouse-Geisser correction. We computed Bayes factors using Bayesian analyses from the R package *BayesFactor* (Morey & Rouder., 2021). Key statistical analyses were determined for each dataset and Bayesian alternatives were computed. Table 2 shows the Bayesian alternative analyses that were employed, as well as providing a citation as to which article each dataset was taken from.

Table 2

Bayesian Alternatives and Dataset Citation

Dataset	Bayesian Alternatives	Research Citation
AgnZI	One-Way ANOVA	Meristo, Strid and Hjelmquist, 2016
DFDwT	Two-Way Repeated Measures ANOVA	Howard, 2016
JcuWB	ANCOVA	Hood et al., 2016
NoMcC	Two-Way Repeated Measures ANOVA	Maister and Tsakiris, 2016
Tbkij	One-Way Repeated Measures ANOVA	Yabe, Dale and Goodale, et al., 2017
UAIUi	One-Way ANOVA, Two-Way ANOVA	Ward, Bear and Scholl,, 2016
UlhiU	One-Sample *t* test, One-Sample *t* test	Wang and Busemeyer, 2016
Wzqlp	Two-Tailed *t* test	Perszyk and Waxman, 2016
Xvfpm	Mixed ANCOVA	Meilinger, Strickrodt and Bülthoff, 2016

Because Bayesian Analyses of Variance introduce proportional error as there is no analytic solution, the 'recompute' function was employed at a default of 1000 iterations for

each ANOVA. If after an initial recomputation the proportional error of any model comparison continued to exceed 5%, this was increased to 5000 iterations, as was the case for five comparisons involving dataset JcuWB. Because Bayesian statistics require the specification of a prior for the parameters of interest, this research defined H_0 for these parameters (α) as equal to 0 (H_0: α = 0). For H_1, we used the default prior of JZS, with a default scale parameter *r* denoted within the BayesFactor package as *medium*, this means the value of the scale parameter *r* is set at 0.5 for the Type 3 ANOVAs, and $\frac{\sqrt{2}}{2}$ for the *t* tests. This was done as we did not have the necessary domain knowledge to specify informed priors for each analysis. This choice of using a default prior is contentious however and is later considered within the context of the wider literature in the discussion section. Within each ANOVA, we used the *top* setting for the argument *whichmodels* to ensure that the resulting Bayes factors are comparable to the *p* values of a typical type 3 ANOVA. As previously discussed, the evidence thresholds were adapted from Wetzels and other author's 2011 paper. As we found only 15% of the Bayes factors from the Bayesian hypothesis alternative testing exceeded 10 or 1/10 (Strong), we excluded thresholds 30 & 1/30 (Very Strong) and 100 & 1/100 (Decisive) from both sequential analysis types.

Procedure

The following steps were conducted in this study; Firstly, the key analyses of each reproduced study were determined, and a Bayes factor was calculated based on the whole dataset using the Bayesian alternatives to these key classical significance tests. To allow for an explicit comparison to Wetzels and other author's 2011 paper, the *p* values and Bayes factors produced from these full datasets was reported and compared. Then, both the fixed order sequential analysis and the replicated random order sequential analysis were conducted for each dataset. For both types of sequential analysis, each dataset was reduced to a pool of participants of the smallest sample size to successfully conduct said Bayesian analysis and a

Bayes factor was then computed. Then subsequent units of participants the size of the smallest unit by which the statistical test was conducted, were added to the pool of participants and another Bayes factor was calculated. This continued until all participants were added to the Sequential analysis, resulting in a sequence of Bayes factors that indicated the trajectory of the Bayes factors throughout the data collection process. For the Fixed order sequential analyses, the order participants were added was the order they were presented in each dataset. In the replicated random order sequential analyses, participants were added in a random order and this Sequential analysis was replicated 100 times, producing 100 trajectories of the Bayes factors. For the fixed order sequential analysis, each sample size at which the Bayes factor exceeded a threshold of interest was reported. For the random order sequential analysis, the Averaged Minimum Sample Size (AMSS) that each threshold was exceeded was reported, alongside the percentage of sequences which produced a Bayes factor that exceeded threshold(s) across all 100 replications of each analysis.

Results

Table 3 shows our findings when comparing the *p* values and Bayes factors produced using both classical and Bayesian hypothesis testing.

Bayes Factors and *p* values Comparison

Table 3

P values and Bayes factor Comparison

Dataset	Effect	*p* value	Bayes factor
AgnZI	*Group*	.222	1.599
DFDwT	*Agency:Effort*	.154	3.485
	Effort	5.750×10^{-4}	1.598
	Agency	1.463×10^{-8}	1.455×10^{-18}
JcuWB	*Gender:Object:Prime*	.863	4.181
	Age:Object	.764	3.579
	Gender:Object	.107	1.041
	Object:Prime	.012	0.101
	Gender:Prime	.129	1.394
	Object	.461	4.126
	Age	.008	0.181
	Gender	.501	3.300
	Prime	.547	5.618
NoMcC	*Congruency:Relationship*	.001	0.123
	Congruency	1.724×10^{-8}	4.883×10^{-12}
	Relationship	.988	4.347
Tbkij	*MeanDifference:Condition*	1.08×10^{-6}	3.142×10^{-5}
UAIUi	*Cue1*	.174	1.354
	Cue2:Diversity	.862	2.780
	Diversity	.77	3.536
	Cue2	.673	3.342
UlhiU	*BadFaces*	.027	1.026
	GoodFaces	.544	9.721
Wzqlp	*PreferenceScore*	.008	0.149
Xvfpm	*CorridorDistance:ECond*	.006	0.002
	CorridorDistance	.002	0.022
	ECond	.834	3.812

Note. Some object names in the effect column have been changed for readability.

Of the 27 classical significance tests conducted, 11 produced *p* values that indicated evidence against H_0 below that of a significance level of .05 (41%). Of these 11 *p* values, four of them produced Decisive evidence against H_0. Three of the Bayes factors for these tests also indicated Decisive evidence for H_1 and one produced Anecdotal evidence for H_0. Five significant *p* values indicated Strong evidence against H_0. For the Bayesian hypothesis testing of these five classical significance tests, one Bayes factor indicated Strong evidence for H_1, two indicated Promising evidence for H_1 and two indicated Substantial evidence for H_1. Two of the 11 significant *p* values indicated Anecdotal evidence against H_0, Bayesian hypothesis

testing produced one Bayes factor that indicated Promising evidence for H_1 and the other indicated Anecdotal evidence for H_0. 16 of the 27 Classical Significance tests produced *p* values that did not indicate notable evidence against H_0 (59%). Of these 16, all Bayesian hypothesis tests produced Bayes factors that indicated support for H_0. One indicated Promising evidence for H_0, 10 indicated Substantial evidence for H_0 and five indicated Anecdotal evidence for H_0.

Both the *p* values and Bayes factors seem to largely covary. All non-significant *p* values had corresponding Bayes factors that indicate support for H_0, and 10 of the 11 *p* values that indicated significant evidence against H_0 produced Bayes factors that indicated support for H_1 (91%). Of the 11 classical significance tests that produced significant *p* values, five of their Bayesian alternative hypothesis tests produced Bayes factors that indicated weaker evidence for H_1 than the *p* values. Three of the 11 Bayesian alternative hypothesis tests indicated a greater degree of support for H_1 than the *p* values (27%). The latter is contradictory to Wetzels et al.'s (2011) research, which found that Bayes factors were more conservative than *p* values, finding that "70% of these experimental effects [*p* values] convey evidence in favor of the alternative hypothesis that is only 'anecdotal' " (Wetzels et al., *p*. 295).

As these findings were contradictory to existing literature, we investigated these three Bayesian analyses further. We suspected that this might be due to an overfit of their priors leading to an overemphasis on the strength of the evidence for H_1. To test this, we adjusted the scale parameter *r* from the default *medium* to *wide* using the *rscaleFixed* argument within the *BayesFactor* package in R and reported these findings below. As per our other Bayesian analyses, we recomputed each analysis by a default of 1000 iterations, or 5000 if the proportional error of items remained above 5%.

Table 4

P values and Rescaled Bayes factors

Effect	*p* value	Original BF	Rescaled BF
Object:Prime	.012	0.101 (1.011×10^{-1})	0.114 (1.148×10^{-1})
CorridorDistance:ECond	.006	0.002 (1.937×10^{-3})	0.004 (3.923×10^{-3})
CorridorDistance	.002	0.022 (2.152×10^{-2})	0.048 (4.792×10^{-2})

When rescaled the Bayes factors become more in line with the *p* values, however this was not to any notable degree. It is therefore unlikely that overfit is the primary reason for these novel findings.

Fixed Order Sequential Analysis with Maximal *n*

After comparing the Bayes factors and *p* values produced from each complete dataset, we then conducted the fixed order sequential analysis, the results of which are described below.

Table 5

Fixed order sequential analysis with Maximal *n*

Dataset	Effect	*N*	BF_{tot}	Smallest Sample Threshold exceeded					
				1/10	1/6	1/3	3	6	10
AgnZI	*Group**	30	1.599				18		
DFDwT	*Agency:Effort**	35	3.485				16		
	*Effort**		1.598				8		
	*Agency**		1.455×10^{-18}	4	2	2			
JcuWB	*Gender:Object:Prime*	60	4.181	6	6	6	30		
	*Age:Object**		3.579				48		
	Gender:Object		1.041			36			
	*Object:Prime**		.101	48	48	42			
	Gender:Prime		1.394						
	Object		4.126			6	42		
	*Age**		.181		54	48			
	*Gender**		3.3				42		
	*Prime**		5.618				42		
NoMcC	*Congruency:Relationship**	19	.123	16	10	8			
	*Congruency**		4.883×10^{-12}	2	2	2			
	*Relationship**		4.347				10		
Tbkij	*MeanDifference:Condition**	19	3.142×10^{-5}	6	6	6			
UAIUi	*Cue1*	12	1.354			8			
	Cue2:Diversity	12	2.780						
	*Diversity**		3.536				10		
	*Cue2**		3.342				10		
UlhiU	*BadFaces**	169	1.026				22	98	
	*GoodFaces**		9.721				22	78	160
Wzqlp	*PreferenceScore**	14	.149		14	10			
Xvfpm	*CorridorDistance:ECond**	24	.002	2	2	2			
	CorridorDistance		.022	24	24	22	6	10	10
	*ECond**		3.812				18		

Note. While contrary to APA guidelines. the font size has been adjusted to size 10 for readability. Asterisks indicate analyses whose initial Bayes factors exceed a threshold indicating support for the same hypothesis as the maximal *n*.

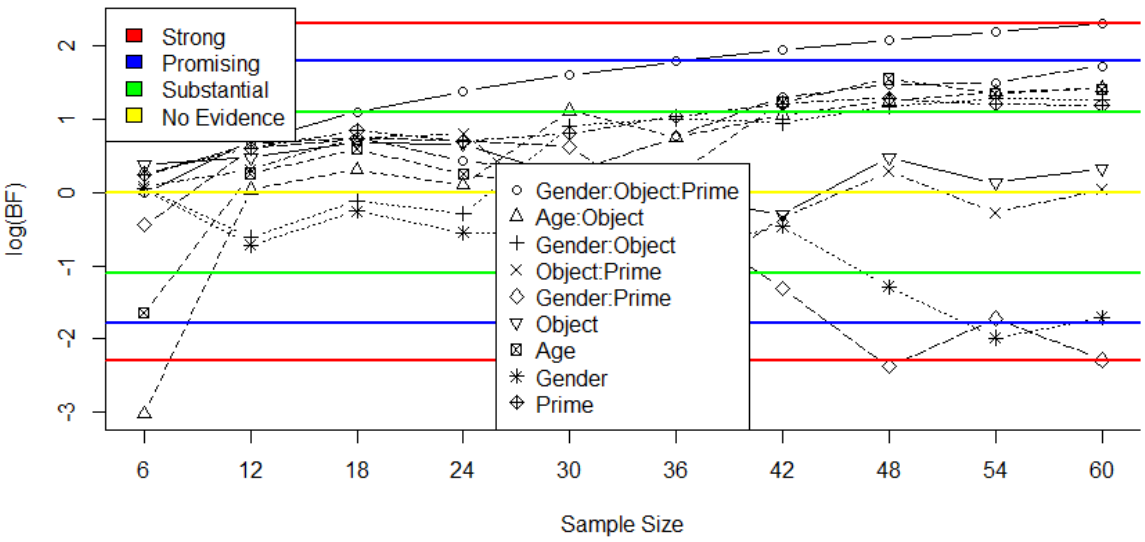
As shown by Table 5, from the 27 fixed order sequential analyses, 25 produced Bayes factors that exceeded at least one threshold prior to reaching the maximal *n*. Of these 25 sequential analyses, 12 exceeded only a single threshold and one exceeded all six thresholds. On average, each analysis produced Bayes factors that exceeded 1.89 thresholds (*SD* = 1.31). In total, the fixed order sequential analyses produced Bayes factors that exceeded 51 thresholds.

One consideration a researcher may have is the effect of predetermined thresholds in optional stopping on efficiency gain. To examine this, we calculated the average efficiency

gain for each fixed order sequential analysis at each threshold band by selecting the smallest initial Bayes factors that exceeded each band. At the Substantial threshold band (3 & 1/3), 25 thresholds were exceeded across all analyses and the average efficiency gain was 49.62%. At the Promising threshold band (6 & 1/6), 12 thresholds were exceeded across all sequential analyses and the average efficiency gain was 24.64%. At the Strong threshold band (10 & 1/10), 9 thresholds were exceeded across all sequential analyses and the average efficiency gain was 19.61%. This indicates that even with at our most stringent thresholds, we could have reduced our sample size on average by ~20% and still identified informative Bayes factors. To maximize efficiency gain, one could pause data collection at the most liberal threshold of Substantial. In practice however, such efficiency gains are not always possible, as one study can produce Bayes factors for multiple effects. This problem is shown below in Figure 2.

Figure 2

Sample Sizes and Bayes factors from JcuWB Dataset, Fixed Order Sequential Analysis



Note. The Log of the Bayes factors was used in this plot

Figure 2 illustrates the nine different factor trajectories derived from the fixed order sequential analysis of the JcuWB Dataset. Each trajectory of Bayes factors varied during the

Sequential analysis, with every effect producing a Bayes factor which exceeded at least two thresholds at some point. This highlights a practical consideration researchers will encounter – sample size considerations regarding one effect cannot be disentangled from another within the same analysis or dataset. Selection of solely the minimum sample size that produces a Bayes factor that exceeds a threshold is therefore not as simple as presented here. This limitation will be explored further within the discussion section.

Furthermore, comparing these initial Bayes factors to the Bayes factor produced by the maximal *n* led to instances of different research conclusions, the degree of the difference in these conclusions based on these fixed order sequential analyses are discussed below. For clarity, the factor produced from the smallest sample size in each sequential analysis that exceeded the Substantial threshold of 3 or 1/3 is referred to as an initial Bayes factor.

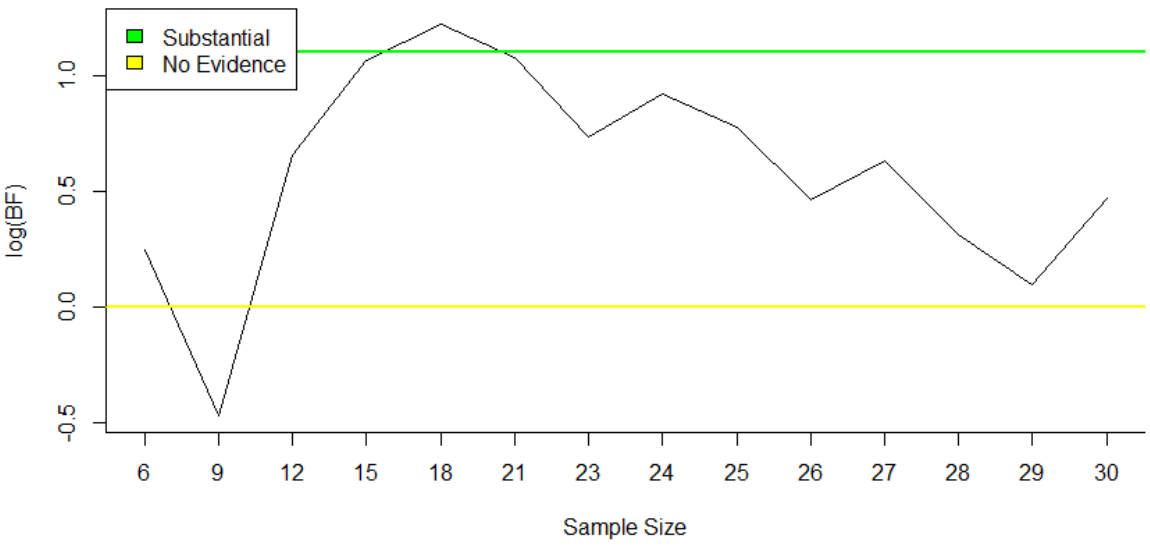
Of the 25 Bayes factors that exceeded at least one threshold, 20 sequential analyses produced initial Bayes factors indicating evidence for the same hypothesis as the Bayes factor from the maximal *n*: These are marked by an asterisk in Table 5. Eight of these 20 analyses produced initial Bayes factors that exceeded the same threshold as the Bayes factor of the maximal *n*. Therefore, in 40% of all sequential analyses, pausing data collection at the initial Bayes factor that exceeded the threshold of Substantial evidence would have produced the same degree of support for the same hypothesis as the maximal *n*. However, 12 analyses produced initial Bayes factors that did not exceed the same threshold as the maximal *n*. One example of this is shown in the analysis of the effect of *Congruency:Relationship* in the NoMcC dataset. The initial Bayes factor exceeded a threshold of 3 with 8 participants, whereas the Bayes factor of the maximal *n* exceeded the threshold of 6.

Four of the 12 analyses produced an initial Bayes factor that indicated a stronger degree of evidence for the hypothesis than the Bayes factor of the maximal *n*. An example of this is illustrated by the sequence of Bayes factors for the effect of *Group* in the AgnZI dataset in

Figure 3 below. The Bayes factors initially indicates Anecdotal evidence for H_0, then reached a global minimum indicating support for H_1 before increasing to indicate Substantial support for H_0 at a sample size of 18 participants. Subsequent Bayes factors indicated weaker evidence for H_0, with the maximal *n* indicating Anecdotal evidence for H_0. If these initial Bayes factors were found during optional stopping, this would lead researchers to conclude that the research overemphasized the support for the hypothesis in question. This is a demonstration of how optional stopping can be overly sensitive to detecting effects. Another example of this is in Figure 1, panel E, where the values peak at the thresholds of 1/10 & 10 which were used in Rouder's (2014) simulation.

Figure 3

Sample Sizes and Bayes factors of Group, AgnZI Dataset, Fixed Order Sequential Analysis



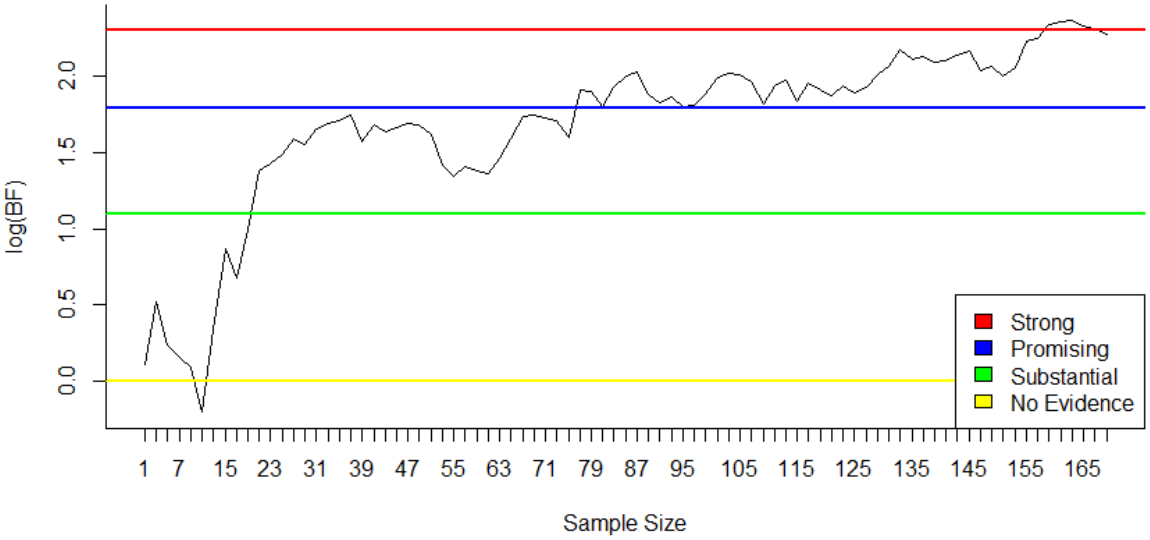
Note. The log of the Bayes factor was used in this plot.

Eight of the 12 analyses produced initial Bayes factors that did not indicate the same degree of support as the Bayes factor of the maximal *n,* indicated a weaker degree of evidence for the same hypothesis than the Bayes factor of the maximal *n*. An example of this is in Figure 4 which shows the sequence of the Bayes factors derived from the effect of

GoodFaces from the dataset UlhiU. Figure 4 shows that the sequence of Bayes factors for the *GoodFaces* initially indicated Anecdotal evidence for H_1. Further participants reversed this trend, with the maximal *n* indicating Strong evidence for H_0. Under optional stopping with the substantive thresholds, researchers would conclude that the evidence for the hypothesis in question is less than it could have been with more participants. These results illustrate a potential drawback with optional stopping, where the aim to reduce the resources invested by the researcher means they fail to detect more convincing evidence for their conclusions.

Figure 4

Sample Sizes and Bayes factors of GoodFaces, UlhiU Dataset, Fixed Order Sequential Analysis



Note. The log of the Bayes factor was used in this plot.

Replicated Random Order Sequential Analysis

To assess our findings from the fixed order sequential analyses, we conducted 27 replicated random order sequential analyses, allowing us to consider our initial findings when compared to hypothetical replications that randomized the order participants were collected. We summarize the results from this below.

Table 6

Replicated random order sequential analysis

Dataset	Effect	BF_{tot}	*N*	Average Minimum Sample Size, Percentage of Total Sequences					
				BF < …				BF > …	
				1/10	1/6	1/3	3	6	10
AgnZI	*Group**	1.599	30				18, 38		
DFDwT	*Agency:Effort*	3.485	35			12, 1	20, 100		
	Effort	1.598				6, 2	12, 54		
	*Agency**	1.455×10^{-18}		4, 100	4, 100	2, 100			
JcuWB	*Gender:Object:Prime*	4.181	60	6, 13	6, 14	6, 24	42, 100		
	Age:Object	3.579			12, 6	6, 21	48, 100		
	Gender:Object	1.041		18, 20	18, 27	24, 46	36, 17		
	Object:Prime	.101		30, 49	36, 100	24, 100	36, 4		
	Gender:Prime	1.394				36, 13	36, 14		
	Object	4.126		12, 3	12. 3	6, 30	30, 100		
	*Age**	.181		36, 28	36, 50	42, 100			
	Gender	3.3				12, 2	48, 100		
	*Prime**	5.618					30, 100	48, 36	
NoMcC	*Congruency:Relationship**	.123	19	16, 32	16, 100	12, 100			
	*Congruency**	4.883×10^{-12}		4, 100	4, 100	2, 100			
	*Relationship**	4.347					12, 100		
Tbkij	*MeanDifference:Condition**	3.142×10^{-5}	19	6, 100	6, 100	4, 100			
UAIUi	*Cue1*	1.354	12			8, 11			
	Cue2:Diversity	2.780		2, 2	2, 2	4, 4			
	Diversity	3.536		2, 1	4, 2	2, 8	10, 100		
	Cue2	3.342			2, 1	4, 3	10, 100		
UlhiU	*BadFaces**	1.026	169	56, 16	54, 33	58, 49	20, 90	56, 43	
	*GoodFaces**	9.721		22, 2	28, 4	22, 7	18, 100	68, 100	142, 71
Wzqlp	*PreferenceScore**	.149	14	10, 11	14, 100	12, 100			
Xvfpm	*CorridorDistance:ECond*	.002	24	18, 100	12, 100	10, 100	2, 79	4, 37	4, 17
	CorridorDistance	.022		18, 100	18, 100	16, 100	4, 89	4, 79	4, 58
	*ECond**	3.812*					18, 100		

Notes. Though this is contradictory to APA guidelines, the font size of this table has been reduced to size 10 for readability.

As illustrated by Table 6, 27 replicated random order sequential analyses were conducted. We calculated the Averaged Minimum Sample Size (AMSS) by averaging the smallest sample size that exceeded a threshold across each of the 100 random orders. We also determined how often the 100 replications contained a Bayes factor that exceeded a particular threshold.

To determine how consistent our fixed order sequential analysis findings were, we calculated the average percentage of total sequences for each of the 25 initial Bayes factor in the fixed order sequential analysis that exceeded the Substantial threshold. We found that on average, the same thresholds exceeded in the fixed order sequential analysis by the initial Bayes factor were exceeded in 83.72% of the replicated random order sequential analyses. This indicated that when reshuffled and replicated, our fixed order sequential analysis findings remained consistent.

Furthermore, 17 replicated random order sequential analyses indicated different sample sizes when comparing the AMSS, to the minimum sample sizes of their fixed order counterparts, the largest difference found was 36 participants and the smallest being 2 participants ($M = 8.71$, $SD = 9.13$).

Discussion

Research Outcomes

Our research aimed to answer two main questions: How do Bayes factors and *p* values from statistical tests other than *t* tests covary (Wetzels et al., 2011), and is Bayesian hypothesis testing suitable within an optional stopping framework, simulated by a maximal *n* Sequential Bayes factor design as per Schönbrodt and Wagenmaker's 2018 paper using real-world data.

After conducting both classical and Bayesian alternatives to *t* tests and Analyses of Variance, our results indicated that *p* values and Bayes factors largely covary when

evaluating the strength of the evidence for a null versus alternative hypothesis as per Wetzels et al.'s (2011) findings. Our findings therefore support the conclusion that Bayes factors are a useful metric by which to conduct hypothesis testing. Further, we found Bayes factors that were more conservative in estimating the strength of evidence in favor of the alternative hypothesis (compared to the null) than *p* values, a solution to what researchers posit is the *p* values tendency to overestimate the evidence against the null hypothesis (Benjamin & Berger, 2019; Jeffreys 1963; Goodman, 1999). However, Wetzels et al.'s observations that Bayes factor were more conservative than *p* values was less consistent when it comes to Bayesian Analyses of Variance, even when ad-hoc testing in which we rescaled the *r* parameter confirmed this was unlikely due to overfit. These notable findings suggest that if Bayes factors are used in Hypothesis testing with Analyses of Variance, the Bayes factors produced might not be more conservative when it comes to supporting the alternative hypothesis stating that there is a main effect/interaction effect. This consideration suggests caution in instances where other measures such as *p* values are not able to assess the strength of the evidence for a hypothesis, as is the case with H_0. While these comparisons partially reaffirmed findings from other researchers, they also produced novel avenues for further exploration, the disproportionate and comparatively small sample when considering other research like this (comparing 25 Analyses of Variance and two *t* tests to the 885 *t* tests compared by Wetzels et al. 2011 most notably) potentially limits the generalizability of these findings. Future research with a greater inventory of research findings would allow for a richer and more nuanced investigation as to the relationship between Bayes factors and *p* values.

We considered whether a maximal *n* sequential analysis using real-world data was a suitable design. We first began by conducting a fixed order sequential analysis to investigate whether, when applied retrospectively, smaller samples than the maximal *n* of each analysis

would have produced compelling evidence, as well as to investigate whether the resulting Bayes factors differed significantly from those at the maximal *n*. Our fixed order sequential analysis found that almost every model comparison produced significant Bayes factors at sequential steps prior to its maximal *n* (93%). As referenced by Schönbrodt and Wagenmakers, "the purpose of a prospective design analysis is to facilitate the design of a study that ensures a sufficiently high probability of detecting if an effect exists" (Schönbrodt & Wagenmakers, 2018, *p*.130). Applying a strict definition, our maximal *n* sequential analyses were successful in detecting an effect prior to the maximal *n*, meaning that if this condition was applied prior to the collection of these datasets, it would have saved researchers resources and time. However, since this is real data, we do not know the underlying truth, so success can only be defined regarding the conclusion drawn at the maximal *n*. Also, it is important to note that a researcher who blindly pauses data collection at the first informative Bayes factor is likely to end up with a very small sample. In most Fixed order sequential analyses, the smallest sample which produced an informative Bayes factor was less than half the size of the maximal *n*. While Schönbrodt and Wagenmakers assert that researcher may set a minimal sample size to be collected, no firm guidelines are proposed. While research in other fields has been conducted to calculate appropriate minimum sample sizes to reach predefined targets in clinical trials (Tan & Machin, 2002; Mayo & Gajewski, 2004) such a topic is less explored within Psychology. Contemporary research into determining sample sizes for Bayesian *t* tests and Welch's test with simulations have been published (Fu, Hoijtink, & Moerbeek, 2021). However, such research is yet to extend to either Analyses of Variance or real-world data. We believe that this is a fertile ground for future research in which Bayesian sample size calculation is coupled with Bayesian sequential analysis, illustrating to researchers in Psychology with real-world examples, the steps they may take to increase their efficiency during data collection, while still producing

results that are “real, reliable, replicable, and hence worthy of academic attention” (Wetzels et al., 2011, *p*. 291).

While our fixed order sequential analysis produced Bayes factors that exceeded 51 thresholds across all 27 analyses, our replicated random order sequential analysis increased this to 84. The replicated random order analysis allowed us to consider how biased our fixed order sequential analysis was to participant order by hypothetically repeating our sequential analysis 100 times with a random order. We found that our fixed order sequential analyses findings were consistent, as a large percentage of the sequences in the replicated random order sequential analyses also found Bayes factors that exceeded the same thresholds as the initial Bayes factors on average 83.72% of the 100 replications of each analysis. Therefore, our replicated random order sequential analysis supports our fixed order sequential analysis findings.

Practical Limitations and Conclusion

While our results are interesting, certain methodological decisions might have negatively impacted various components of our research. Within this Discussion section we begin by considering the relatively minor decisions (choices on key reproducible analyses, differences in threshold size between *p* values and Bayes factors, mostly statistical tests producing non-significant *p* values). We then discuss the methodological decisions that may have a more serious impact on the generalizability of our findings (our choice of priors and the utility of optional stopping with Bayesian Hypothesis Testing).

When selecting key reproducible analyses, we did not include a linear mixed-effect analysis from the NoMcC dataset due to it being produced using SPSS Version 24. As a result, not all the statistical analyses that we categorized as key were considered. Future research that includes reproducing existing analyses might consider using multiple statistical programs to avoid such a problem.

Furthermore, this research used thresholds adapted from Wasserman and other author's 2011 research for classifying both the *p* values and Bayes factors. As a result, however, *p* values were placed in fewer threshold bins than Bayes factors, meaning that *p* values from .01 and smaller covered one threshold bin, unlike their Bayes factor counterparts which covered four. The use of threshold in general however is somewhat debatable. While a Bayes factor threshold of 3, 1/3 was deemed useful due to its rough approximation of the *p* value significance level of .05 (Jeffrey, 1963). Recent efforts made by the scientific community to propose changing the significance level of a *p* value to .005 (Benjamin et al., 2019) illustrate the degree to which even this threshold is arbitrary. While thresholds allow clear classification of Bayes factors, it should be emphasized that exceeding a threshold is not the only metric by which to assess Bayes factors.

Different statistical tests required different numbers of minimum participants. As a result, different sequential tests began with a minimum pool of participants of varying sizes. To mimic real decisions researchers may make during data collection, we added an equal number of participants to each group during each sequential step, as Analyses of Variance with unequal group sizes can lead to pronounced Type I error rates (Troncoso Skidmore & Thompson, 2013). As such however, different datasets of similarly sizes had significantly more or less sequential steps prior to the maximal *n*. This was a limiting factor in capturing nuance when evaluating and comparing the results of the Sequential analysis across datasets. While a position we still believe to be valid, it did limit the number of BF we collected within every Sequential analysis.

As with all maximal *n* analyses, we are reliant on assuming that the maximal *n* is more indicative of the 'truth' than the smaller sequential samples. While this is intuitively likely, as we have demonstrated in Figures 2-4, Bayes factors can shift in not only the strength of their support for a hypothesis, but as support for either H_1 or H_0. It is not

unthinkable therefore to assume that one or more of these maximal *n*'s might not be indicative of the degree of a Bayes factor calculated using more data. Such a consideration was apparent to us, and while comparisons were made between the BF of the maximal *n* to those produced from each Sequential step, we avoided considering one BF or another 'false'.

For utility, we selected default priors used by the *BayesFactor* package. This was to minimize the number of assumptions we made prior to analysis, as we did not feel that we have the required domain knowledge to create more informed priors. Such an argument stems from a particular view of the philosophical framework by which to consider the relationship between Bayesian Statistics and probability, so called *Objective Bayesianism* (Joyce, 2004). Objective Bayesians assert that there is a way by which to set priors such that they correctly represent uncertainty. Other Bayesians would disagree with this approach however, arguing that priors represent the subjective degrees of belief, or *credences,* by the researcher. This perspective is referred to as *Subjectivist Bayesianism* (Joyce, 2011). While Subjectivist Bayesians would assert that any prior is sufficient, they would also posit that this is due to it representing the agent's personal degree of belief (de Finetti, 1937). However, researchers have argued that by de Finetti's definition of personal belief as "a willingness to bet at small stakes, at the odds given by the prior.", Bayesians would not agree to such a definition in practice (De Heide & Grünwald, 2021, *p*. 808). When investigating the impact of default priors on optional stopping, De Heide and Grünwald (2021) took issue with the use of default priors within optional stopping, namely that by Rouder's 2014 definition of how the Bayesian method can interpret results used by optional stopping, which they referred to as *prior calibration*, are of limited relevance or undefined. De Heide and Grünwald (2021) found that while optional stopping within a Bayesian framework is interpretable as found by Rouder's (2014) simulations, they found that sampling from the parameter prior of H_1 is not practical, as this parameter would reflect a single fixed population value. De Heide and Grünwald

(2021) also conducted two separate replications of Rouder's (2014) simulations, one where the σ^2 is unknown and treated as a nuisance parameter and again when each of the 40,000 experiments were sampled until the posterior odds were at least 10-to-1. In both conditions, they found that Rouder's simulation no longer demonstrated a calibration between the observed posterior odds and the nominal posterior odds. Such findings they argued, indicated that by this *strong calibration*, the accuracy of Bayes factors derived from optional stopping which employs default parameter priors is questionable. De Heide and Grünwald advised researchers who employ optional stopping to use robust analyses referred to collectively as *safe tests* to reduce the likelihood of Type I errors and to reach the threshold of strong calibration (Grünwald, De Heide, & Koolen, 2020). As the employment of default parameter priors seems to be problematic in instances, future research should conduct robust analyses on the use of default priors would clearly mitigate Type I error which may have impacted the findings of this research.

In conclusion, while certain methodological decisions were taken which to varying degrees impacted our experiment and should therefore influence the reader's evaluation of our findings, this research produced novel findings which further evaluated the use of Bayesian hypothesis testing with real-world data across a larger range of statistical tests than previously investigated. This research also demonstrated the utility of maximal *n* sequential analyses with Bayes factors using real-world data and obtained findings that supported conclusions reached by Schönbrodt and Wagenmaker's (2018) research, but also produced results which were contradictory to our expectations and warranted further study. In our mind Sequential analysis can be efficient as it clearly produces informative Bayes factors that would reduce the number of participants a researcher would need to gather. Further, when conducting replicated random order sequential analysis during data collection, this can both reduce the bias which the participant order can introduce in data collection, by reshuffling

existing participants and allowing researchers to consider informative Bayes factors in the aggregate, it can also allow exclusion criteria to be considered and allow a more nuanced perspective on the data being collected. Further, we believe that with future research that attempts to calculate appropriate minimum sample sizes for Bayesian hypothesis testing using Analyses of Variance and evaluated using real world data, we believe that optional stopping would be an efficient way for researchers to collect data.

References

Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53*, 370-418. http://dx.doi.org/10.1093/biomet/45.3-4.296

Benjamin, D. J., & Berger, J. O. (2019). Three recommendations for improving the use of *p* values. *The American Statistician, 73*(sup1), 186-191. http://dx.doi.org/10.1080/00031305.2018.1543135

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

De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré, 7*(1), 1-68.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*(3), 274-290. http://dx.doi.org/10.1177/1745691611406920

Fu, Q., Hoijtink, H., & Moerbeek, M. (2021). Sample-size determination for the Bayesian t test and Welch's test using the approximate adjusted fractional Bayes factor. *Behavior Research Methods, 53*(1), 139-152. http://dx.doi.org/10.3758/s13428-020-01408-1

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall/CRC. http://dx.doi.org/10.1201/b16018

Greenland, S., & Poole, C. (2013). Living with P values: Resurrecting a bayesian perspective on frequentist statistics. *Epidemiology, 24*(1), 62-68. http://dx.doi.org/10.1097/EDE.0b013e3182785741

Grünwald, P., De Heide, R., & Koolen, W. M. (2020, February). *Safe testing.* 2020 Information Theory and Applications Workshop (ITA), 2020, 1-54. http://dx.doi.org/10.1109/ITA50056.2020.9244948

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., ... & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. Retrieved from https://osf.io/preprints/bitss/39cfb/. http://dx.doi.org/10.1098/rsos.180448

Hood, B., Weltzien, S., Marsh, L., & Kanngiesser, P. (2016). Picture yourself: SelF focus and the endowment effect in preschool children. *Cognition, 152*, 70-77. http://dx.doi.org/10.1016/j.cognition.2016.03.019

Howard, G. R. (2016). *We can't teach what we don't know: White teachers, multiracial schools.* Teachers College Press.

Jeffreys, H. (1963). *Theory of Probability* (3rd ed.). Oxford University Press. http://dx.doi.org/10.1063/1.3050814

Joyce, J. M. (2004). Bayesianism. In A. R. Mele and P. Rawlings (Eds.), *The Oxford Handbook of Rationality* (pp. 132-155), Oxford University Press. http://dx.doi.org/10.1093/0195145399.003.0008

Joyce, J. M. (2011). The development of subjective Bayesianism. In D. M. Gabbay, S. Hartmann, and J. Woods (Eds.), *Handbook of the History of Logic, Vol 10: Inductive Logic* (pp. 415-475). Elsevier. http://dx.doi.org/10.1016/B978-0-444-52936-7.50012-4

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773-795. http://dx.doi.org/10.1080/01621459.1995.10476572

Maister, L., & Tsakiris, M. (2016). Intimate imitation: Automatic motor imitation in romantic relationships. *Cognition, 152*, 108-113. http://dx.doi.org/10.1016/j.cognition.2016.03.018

Marszalek, J. M., Barber, C., Kohlhart, J., & Cooper, B. H. (2011). Sample size in psychological research over the past 30 years. *Perceptual and motor skills, 112*(2), 331-348. http://dx.doi.org/10.2466/03.11.PMS.112.2.331-348

Mayo, M. S., & Gajewski, B. J. (2004). Bayesian sample size calculations in phase II clinical trials using informative conjugate priors. *Controlled Clinical Trials, 25*(2), 157-167. http://dx.doi.org/10.1016/j.cct.2003.11.006

Meilinger, T., Strickrodt, M., & Bülthoff, H. H. (2016). Qualitative differences in memory for vista and environmental spaces are caused by opaque borders, not movement or successive presentation. *Cognition, 155*, 77-95. http://dx.doi.org/10.1016/j.cognition.2016.06.003

Meristo, M., Strid, K., & Hjelmquist, E. (2016). Early conversational environment enables spontaneous belief attribution in deaf children. *Cognition, 157*, 139-145. http://dx.doi.org/10.1016/j.cognition.2016.08.023

Morey, R. D., & Rouder, J. N. (2021). *BayesFactor: Computation of Bayes Factors for Common Designs. R package version 0.9.12-4.3.* https://CRAN.R-project.org/package=BayesFactor

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301. http://dx.doi.org/10.1037/1082-989X.5.2.241

Perszyk, D. R., & Waxman, S. R. (2016). Listening to the calls of the wild: The role of experience in linking language and cognition in young infants. *Cognition, 153*, 175-181. http://dx.doi.org/10.1016/j.cognition.2016.05.004

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301-308. http://dx.doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356-374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16(*2), 225-237. http://dx.doi.org/10.3758/PBR.16.2.225

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review, 21*(2), 283-300. http://dx.doi.org/10.3758/s13423-013-0518-9

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*(1), 128-142. http://dx.doi.org/10.3758/s13423-017-1230-y

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. http://dx.doi.org/10.1177/0956797611417632

Tan, S. B., & Machin, D. (2002). Bayesian two-stage designs for phase II clinical trials. *Statistics in Medicine, 21*(14), 1991-2012. http://dx.doi.org/10.1002/sim.1176

Troncoso Skidmore, S., & Thompson, B. (2013). Bias and precision of some classical ANOVA effect sizes when assumptions are violated. *Behavior Research Methods, 45*(2), 536-546. http://dx.doi.org/10.3758/s13428-012-0257-2

Wang, Z., & Busemeyer, J. R. (2016). Interference effects of categorization on decision making. *Cognition, 150*, 133-149. http://dx.doi.org/10.1016/j.cognition.2016.01.019

Ward, E. J., Bear, A., & Scholl, B. J. (2016). Can you perceive ensembles without perceiving individuals? The role of statistical perception in determining whether awareness overflows access. *Cognition, 152*, 78-86. http://dx.doi.org/10.1016/j.cognition.2016.01.010

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6*(3), 291-298. http://dx.doi.org/10.1177/1745691611406923

Yabe, Y., Dave, H., & Goodale, M. A. (2017). Temporal distortion in the perception of actions and events. *Cognition, 158*, 1-9. http://dx.doi.org/10.1016/j.cognition.2016.10.009

Yu, E.C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review, 21*(2), 268-282 (2014). http://dx.doi.org/10.3758/s13423-013-0495-z

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadística y de Investigación Operativa, 31*(1), 585-603. http://dx.doi.org/10.1007/BF02888369