ORIGINAL RESEARCH PAPER

Can we disregard the whole model? Omnibus non-inferiority testing for R^2 in multivariable linear regression and $\hat{\eta}^2$ in ANOVA.

Harlan Campbell[a], Daniël Lakens[b]

[a]University of British Columbia, Department of Statistics

[b]Eindhoven University of Technology

ARTICLE HISTORY

Compiled January 15, 2020

Abstract

Determining a lack of association between an outcome variable and a number of different explanatory variables is frequently necessary in order to disregard a proposed model (i.e., to confirm the lack of a meaningful association between an outcome and predictors). Despite this, the literature rarely offers information about, or technical recommendations concerning, the appropriate statistical methodology to be used to accomplish this task. This paper introduces non-inferiority tests for ANOVA and linear regression analyses, that correspond to the standard widely used F-test for $\hat{\eta}^2$ and R^2, respectively. A simulation study is conducted to examine the type I error rates and statistical power of the tests, and a comparison is made with an alternative Bayesian testing approach. The results indicate that the proposed non-inferiority test is a potentially useful tool for "testing the null."

KEYWORDS

equivalence testing, non-inferiority testing, ANOVA, F-test, linear regression

The data that support the findings of this study are openly available in the OSF repository "Can we disregard the whole model?". at http://doi.org/10.17605/OSF.IO/3Q2VH, reference number 3Q2VH.

CONTACT Harlan Campbell. Email: harlan.campbell@stat.ubc.ca

1. Introduction

All too often, researchers will conclude that the effect of an explanatory variable, X, on an outcome variable, Y, is absent when a null-hypothesis significance test (NHST) yields a non-significant p-value (e.g., when the p-value > 0.05). Unfortunately, such an argument is logically flawed. As the saying goes, "absence of evidence is not evidence of absence" (Hartung et al., 1983; Altman and Bland, 1995). Indeed, a non-significant result can simply be due to insufficient power, and while a null-hypothesis significance test can provide evidence to *reject* the null hypothesis, it cannot provide evidence *in favour* of the null (Quertemont, 2011).

Let θ be the parameter of interest representing the true association between X and Y in the population of interest. The equivalence/non-inferiority test reverses the question that is asked in a NHST. Instead of asking whether we can reject the null hypothesis, e.g., $H_0 : \theta = 0$, an equivalence test examines whether the magnitude of θ is at all meaningful: *Can we reject an association between X and Y as large or larger than our smallest effect size of interest, Δ?* The null hypothesis for an equivalence test is therefore defined as $H_0 : |\theta| \geq \Delta$. Or for the one-sided non-inferiority test, the null hypothesis is $H_0 : \theta \geq \Delta$. Note that researchers must decide which effect size is considered meaningful or relevant (Lakens et al., 2018), and define Δ accordingly, prior to observing any data; see Campbell and Gustafson (2018b) for details.

In a standard multivariable linear regression model, or a standard ANOVA analysis, the variability of the outcome variable, Y, is attributed to multiple different explanatory variables, $X_1, X_2, ..., X_p$. Researchers will typically report the linear regression model's R^2 statistic, or the $\hat{\eta}^2$ in the ANOVA context, to estimate the proportion of variance in the observed data that is explained by the model. To determine whether or not we can reject the hypothesis that the variance attributed to the explanatory variables is equal to zero, one typically calculates an F-statistic and tests whether the "null model" (i.e., the intercept only model) can be rejected in favour of the "full model" (i.e., the model with all explanatory variables included). However, in this multivariate setting, while rejecting the "null model" is rather simple, concluding

in favour of the "null model" is less obvious.

If the explanatory variables are not statistically significant, can we simply disregard the full model? We certainly shouldn't pick and choose which variables to include in the model based on their significance (it is well known that due to model selection bias, most step-wise variable selection schemes are to be avoided; see Hurvich and Tsai (1990)). How can we formally test whether the proportion of variance attributable to the full set of explanatory variables is too small to be considered meaningful? In this article, we introduce a non-inferiority test to reject effect sizes that are as large or larger than the smallest effect size of interest as estimated by either the R^2 statistic or the $\hat{\eta}^2$ statistic.

In Section 2, we introduce a non-inferiority test for the coefficient of determination parameter in a linear regression context. We show how to define hypotheses and calculate a valid p-value for this test based on the R^2 statistic. We then consider how this frequentist test compares to a Bayesian testing scheme based on Bayes Factors, and conduct a small simulation study to better understand the test's operating characteristics. In Section 3, we illustrate the use of this test with data from a recent study about the absence of the Hawthorne effect. In Section 4, we present the analogous non-inferiority test for the η^2 parameter in an ANOVA. We also provide a modified version of this test that allows for the possibility that the variance across groups is unequal.

2. A non-inferiority test for the coefficient of determination parameter

The coefficient of determination, commonly known as R^2, is a sample statistic used in almost all fields of research. In a linear regression model, the R^2 is equal to the square of the Pearson correlation coefficient between the observed and predicted outcomes (Nagelkerke et al., 1991; Zou et al., 2003). Despite the R^2 statistic's ubiquitous use, its corresponding population parameter, which we will denote as P^2, as in Cramer (1987), is rarely discussed. When considered, it is sometimes known as the "parent multiple correlation coefficient" (Barten, 1962) or the "population proportion of vari-

ance accounted for" (Kelley et al., 2007). See Cramer (1987) for a technical discussion.

While confidence intervals for P^2 have been studied by many researchers (e.g., Ohtani and Tanizaki (2004), Ohtani (2000), Dudgeon (2017)), there has been no consideration (as far as we know) of a non-inferiority test for P^2. In this section we will derive such a test and investigate how it compares to a popular Bayesian alternative (Rouder and Morey, 2012). Before we continue, let us define some notation. All technical details are presented in the Appendix. Let:

- N, be the number of observations in the observed data;
- K, be the number of explanatory variables in the linear regression model;
- y_i, be the observed value of random variable Y for the ith subject;
- x_{ji}, be the observed value of fixed covariate X_j, for the ith subject, for k in $1, ..., K$; and
- X, be the N by $K + 1$ covariate matrix (with a column of 1s for the intercept; we use the notation $X_{i,\cdot}$ to refer to all $K + 1$ values corresponding to the ith subject).

We operate under the standard linear regression assumption that observations in the data are independent and normally distributed with:

$$Y_i \sim \ Normal(X_{i,\cdot}^T \beta, \sigma^2), \qquad \forall\, i = 1, ..., N; \tag{1}$$

where β is a parameter vector of regression coefficients, and σ^2 is the population variance. The parameter P^2 represents the proportion of total variance in the population that can be accounted for by knowing the covariates, i.e., by knowing X. As such, P^2 is entirely dependent on the particular design matrix X, and we have that:

$$P^2 = \frac{\sigma_{XY}^T \Sigma_X^{-1} \sigma_{XY}}{\sigma_Y^2}, \tag{2}$$

where σ_Y^2 is the unconditional variance of Y, (note that: $\sigma_Y^2 \geq \sigma^2$); σ_{XY} is the vector of population covariances between the K different X variables and Y; and Σ_X is the population covariance matrix of the K different X variables. The R^2 statistic estimates the parameter P^2 from the observed data. See Kelley et al. (2007) for a complete derivation of equation (2).

A standard NHST asks whether we can reject the null hypothesis that P^2 is equal to zero ($H_0 : P^2 = 0$). The p-value for this NHST is calculated as:

$$p - \text{value} = 1 - p_f(F; K, N - K - 1, 0), \tag{3}$$

where $p_f(\cdot \quad ; df_1, df_2, ncp)$ is the cdf of the non-central F-distribution with df_1 and df_2 degrees of freedom, and non-centrality parameter ncp (note that $ncp = 0$ corresponds to the *central* F-distribution); and where:

$$F = \frac{R^2/K}{(1 - R^2)/(N - K - 1)}. \tag{4}$$

One can calculate the above p-value in `R` with the following code:

```
pval = pf(Fstat, df1 = K, df2 = N-K-1, lower.tail = FALSE).
```

A non-inferiority test for P^2 is asking a different question: *can we reject the hypothesis that the total proportion of variance in Y attributable to X is greater than or equal to Δ?* Formally, the hypotheses for the non-inferiority test are:

$H_0 : 1 > P^2 \geq \Delta$,

$H_1 : 0 \leq P^2 < \Delta$.

The p-value for this non-inferiority test is obtained by inverting the one-sided CI for P^2 (see Appendix for details), and can be calculated as:

$$p - \text{value} = p_f \left(F; K, N - K - 1, \frac{N\Delta}{(1 - \Delta)} \right). \tag{5}$$

Note that one can calculate the above p-value in `R` with the following code:

```
pval = pf(Fstat, df1=K, df2=N-K-1, ncp=(N*Delta)/(1-Delta), lower.tail=TRUE).
```

Under the assumption that the true value of $P^2 = 0$, for given values of N, K, and Δ, a simple analytic formula provides a reasonable approximation of the non-inferiority test's statistical power:

$$power = Pr(\text{reject } H_0 | P^2 = 0) = p_f(F^*; K, N - K - 1, 0), \tag{6}$$

where F^* is equal to the $(1 - \alpha)\%$ critical value of a non-central F-distribution with $df_1 = K$ and $df_2 = N - K - 1$ degrees of freedom, and non-centrality parameter $ncp = (N\Delta)/(1 - \Delta)$.

Note that one can calculate the above power estimate in `R` with the following code:

```
Fstatstar = qf(alpha, df1=K, df2=N-K-1, ncp=(N*Delta)/(1-Delta), lower.tail=TRUE)

power = pf(Fstatstar,df1=K,df2=N-K-1,lower.tail=TRUE).
```

It is important to remember that the above tests make two important assumptions about the data:

- The data are independent and normally distributed as described in equation (1).
- The values for X in the observed data are fixed and their distribution in the sample is equal (or representative) to their distribution in population of interest.

The sampling distribution of R^2 can be quite different when regressor variables are random; see Gatsonis and Sampson (1989).

Ideally, a researcher uses the non-inferiority test to examine a preregistered hypothesis concerning the absence of a meaningful effect. However, in practice, one might first conduct a NHST (i.e., calculate a p-value, p_1, using equation (3)) and only proceed to conduct the non-inferiority test (i.e., calculate a second p-value, p_2, using equation (5)) if the NHST fails to reject the null. Such a two-stage sequential testing scheme has recently been put forward by Campbell and Gustafson (2018a) under the name of "conditional equivalence testing" (CET). Under the proposed CET scheme, if the first p-value, p_1, is less than the type 1 error α-threshold (e.g., if $p_1 < 0.05$), one concludes with a "positive" finding: P^2 is significantly greater than 0. On the other hand, if the first p-value, p_1, is greater than α and the second p-value, p_2, is smaller than α (e.g., if $p_1 \geq 0.05$ and $p_2 < 0.05$), one concludes with a "negative" finding: there is evidence of a statistically significant non-inferiority, i.e., P^2 is at most negligible. If both p-values are large, the result is inconclusive: there are insufficient data to support either finding.

In this paper, we are not advocating for (or against) CET but simply use it to facilitate a comparison with Bayes Factor testing (which also categorizes outcomes as either positive, negative or inconclusive). Other possible testing strategies available to researchers include: (1) performing only an equivalence test, (2) performing both an equivalence test and a NHST (acknowledging the possibility there is a non-zero, but trivial, effect), and (3) performing a NHST if and only if the equivalence test is not significant. As long as these procedures are chosen and performed transparently (e.g., in a preregistered study) there are scenarios for which all these options can be defended.

2.1. Comparison to a Bayesian alternative

For linear regression models, based on the work of Liang et al. (2008), Rouder and Morey (2012) propose using Bayes Factors (BFs) to determine whether the data, as

summarized by the R^2 statistic, support the null or the alternative model. This is a common approach used in psychology studies (e.g., see most recently Hättenschwiler et al. (2019)). Here we refer to the null model ("Model 0") and alternative (full) model ("Model 1") as:

$$\text{Model } 0 : Y_i \sim \quad Normal(\beta_0, \sigma^2), \qquad \forall i = 1, ..., N; \tag{7}$$

$$\text{Model } 1 : Y_i \sim \quad Normal(X_{i,\cdot}^T \beta, \sigma^2), \qquad \forall i = 1, ..., N; \tag{8}$$

where β_0 is the overall mean of Y (i.e., the intercept).

We define the Bayes Factor, BF_{10}, as the probability of the data under the alternative model relative to the probability of the data under the null:

$$BF_{10} = \frac{Pr(Data \,|\, Model\ 1)}{Pr(Data \,|\, Model\ 0)}, \tag{9}$$

with the "10" subscript indicating that the full model (i.e., "Model 1") is being compared to the null model (i.e., "Model 0"). The BF can be easily interpreted. For example, a BF_{10} equal to 0.10 indicates that the null model is ten times more likely than the full model.

Bayesian methods require one to define appropriate prior distributions for all model parameters. Rouder and Morey (2012) suggest using "objective priors" for linear regression models and explain in detail how one may implement this approach. We will not discuss the issue of prior specification in detail, and instead point interested readers to Consonni et al. (2008) who provide an in-depth overview of how to specify prior distributions for linear models.

Using the BayesFactor package in `R` (Morey et al., 2015) with the function `linearReg.R2stat()`, one can easily obtain a BF corresponding to given values for

R^2, N, and K. (Alternatively, see also the `baymedr` package in `R` (Linde and van Ravenzwaaij, 2019)). Since we can also calculate frequentist p-values corresponding to given values for R^2, N, and K (see equations (3) and (5)), a comparison between the frequentist and Bayesian approaches is relatively straightforward.

For three different values of K (=1, 5, 12) and a broad range of values of N (76 values from 30 to 1,000), we calculated the R^2 values corresponding to a BF_{10} of 1/3 ("moderate evidence" in favour of the null model (Jeffreys, 1961)) and of 3 ("moderate evidence" in favour of the full model). We then proceeded to calculate the corresponding frequentist p-values for NHST and non-inferiority testing for the (R^2, K, N) combinations. Note that all priors required for calculating the BF were set by simply selecting the default settings of the `linearReg.R2stat()` function (with rscale = "medium"), whereby a noninformative Jeffreys prior is placed on the variance of the normal population, while a scaled Cauchy prior is placed on the standardized effect size; see Morey et al. (2015).

The results are plotted in Figure 1. The left-hand column plots the conclusions reached by frequentist testing (i.e., the CET sequential testing scheme). For all calculations, we defined $\alpha = 0.05$ and $\Delta = 0.10$. The right-hand column plots the conclusions reached based on the Bayes Factor with a threshold of 3.

Each conclusion corresponds to a different colour in the plot: *green* indicates a positive finding (evidence in favour of the full model); *red* indicates a negative finding (evidence in favour of the null model); and *yellow* indicates an inconclusive finding (insufficient evidence to support either model). Note that we have also included a third colour, light-green. For the frequentist testing scheme, light-green indicates a scenario where both the NHST p-value and the non-inferiority test p-value are less than $\alpha = 0.05$. The tests reveal that the observed effect size is both statistically significant (i.e., we reject $H_0 : P^2 = 0$) and statistically smaller than the effect size of interest (i.e., we also reject $H_0 : P^2 \geq \Delta$). In these situations, one could conclude that, while P^2 is significantly greater than zero, it is likely to be practically insignificant (i.e., a real effect of a negligible magnitude).

Three observations merit comment:

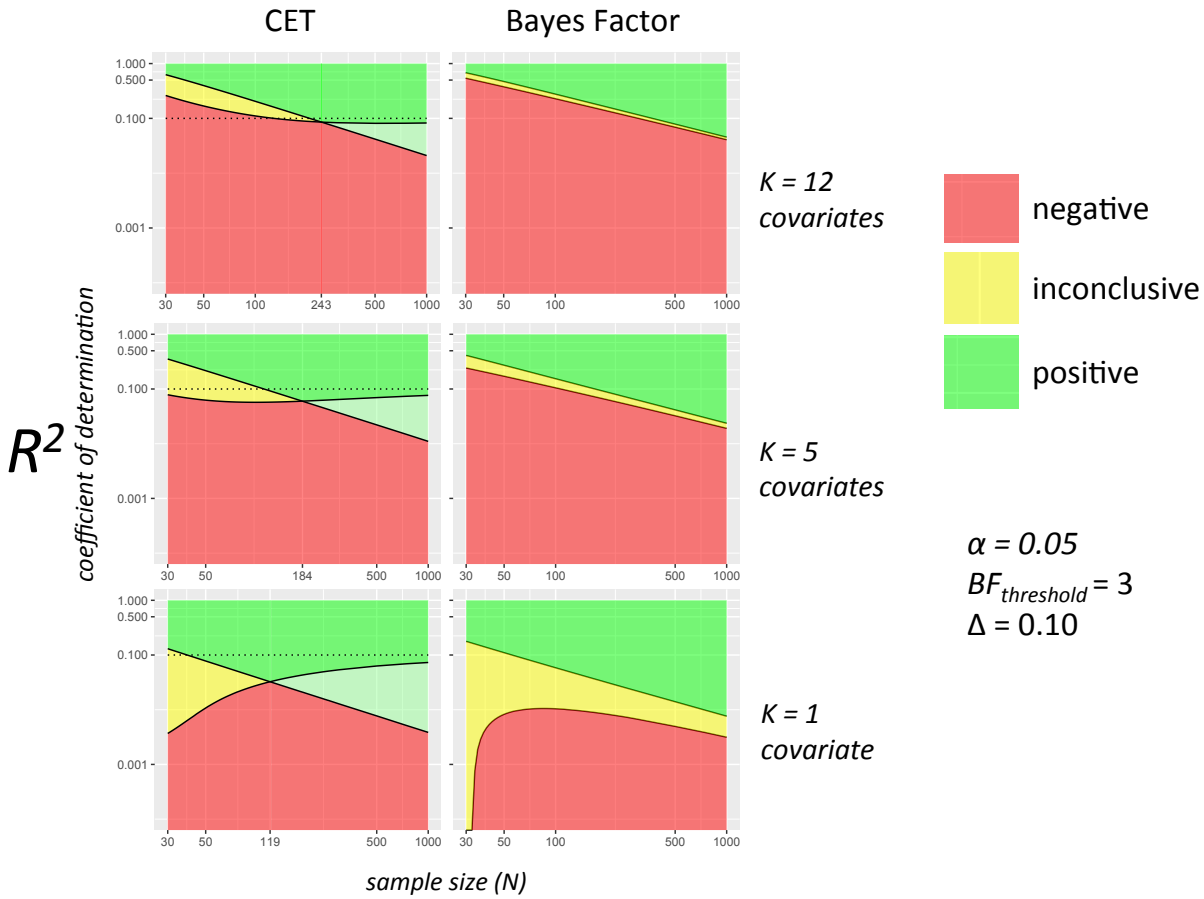

Figure 1. Colours indicate the conclusions corresponding to varying levels of R^2 and N (red="negative"; yellow = "inconclusive"; green="positive"). Left panels shows the frequentist testing scheme with NHST and non-inferiority test ($\Delta = 0.10$) and right panels show Bayesian testing scheme with a threshold for the BF of 3. The significance threshold for frequentist tests is $\alpha = 0.05$. Both vertical-axis (R^2) and horizontal-axis (N) are on logarithmic scales. Note that the "light-green" colour corresponds to scenarios for which both the NHST and the non-inferiority p-values are less than $\alpha = 0.05$. One could describe the effect in these cases as "significant yet not meaningful."

(1) When testing with Bayes Factors, there will always exist a combination of values of R^2 and N that corresponds to an inconclusive result. This is not the case for frequentist testing: the probability of obtaining an inconclusive finding will decrease with increasing N, and at a certain point, will be zero. For example, with $K = 5$ and any $N > 184$, it is impossible to obtain an inconclusive finding regardless of the observed R^2.

(2) For $K = 1$ covariate, with $N < 30$, it is practically impossible to obtain a negative conclusion with the Bayesian approach, and only possible with the frequentist approach (for the equivalence bound of $\Delta = 0.10$), if the R^2 is very very small ($\approx< 0.001$).

(3) For $K = 12$ covariates, with $N < 50$, the frequentist testing scheme obtains a negative conclusion in situations when $R^2 > \Delta$. This may seem rather odd but can be explained by the fact that R^2 is "seriously biased upward in small samples" (Cramer, 1987).

Based on this comparison of BFs and frequentist tests, we can speculate that, given the same data, both approaches will often provide one with the same overall conclusion. In Section 2.3, we investigate this further by means of a simulation study.

2.2. Simulation study 1

We conducted a simple simulation study in order to better understand the operating characteristics of the non-inferiority test and to confirm that the test has correct type 1 error rates. We simulated data for each of twenty-four scenarios, one for each combination of the following parameters:

- one of four sample sizes: $N = 60$, $N = 180$, $N = 540$, or, $N = 1,000$;
- one of two designs with $K = 2$, or $K = 4$ binary covariates, (with an orthogonal, balanced design), and with $\beta = (0.0, 0.2, 0.3)$ or $\beta = (0.0, 0.2, 0.2, -0.1, -0.2)$; and
- one of three variances: $\sigma^2 = 0.4$,$\sigma^2 = 0.5$, or $\sigma^2 = 1.0$.

Depending on the particular values of K and σ^2, the true coefficient of determination

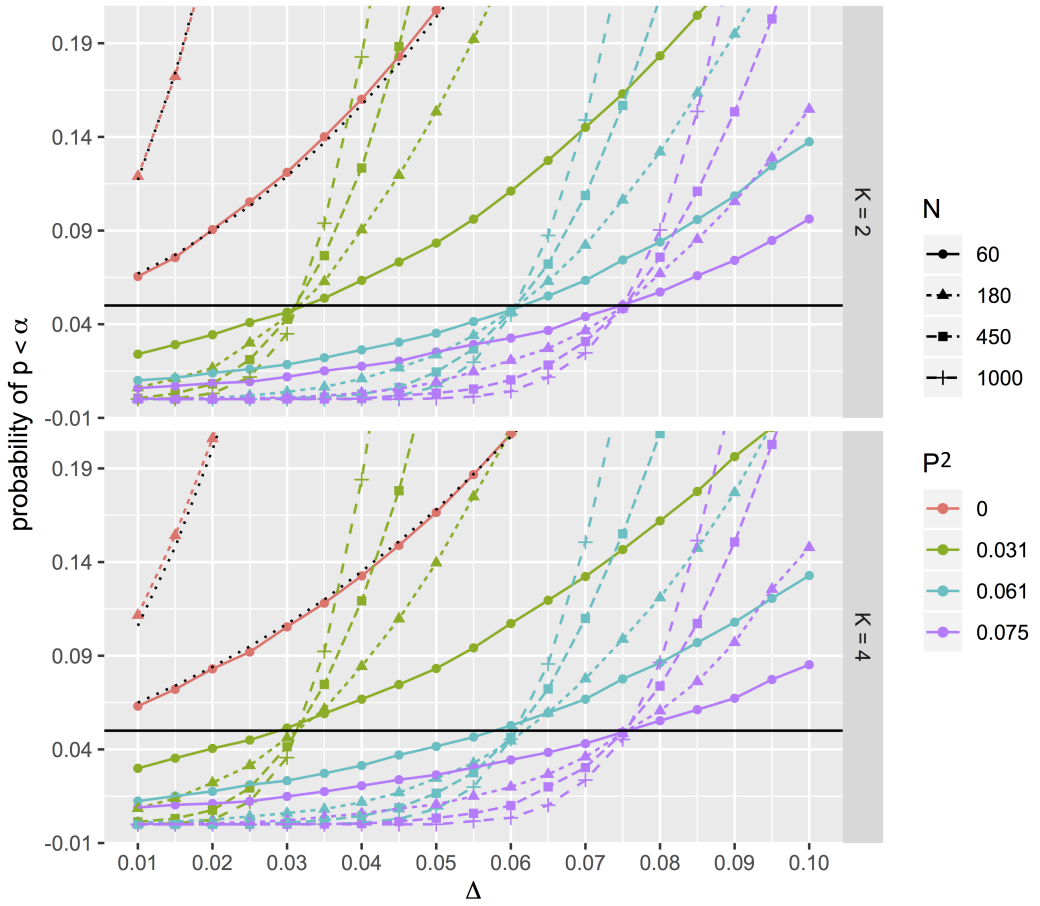

Figure 2. Simulation study results. Upper panel shows results for $K = 2$; lower panel shows results for $K = 4$. Both plots are presented with a restricted vertical-axis to better show the type 1 error rates. The solid horizontal black line indicates the desired type 1 error of $\alpha = 0.05$. The dotted black curves plot numbers calculated using equation (6) for estimating power. For each of thirty-two configurations, we simulated 50,000 unique datasets and calculated a non-inferiority p-value with each of 19 different values of Δ (ranging from 0.01 to 0.10).

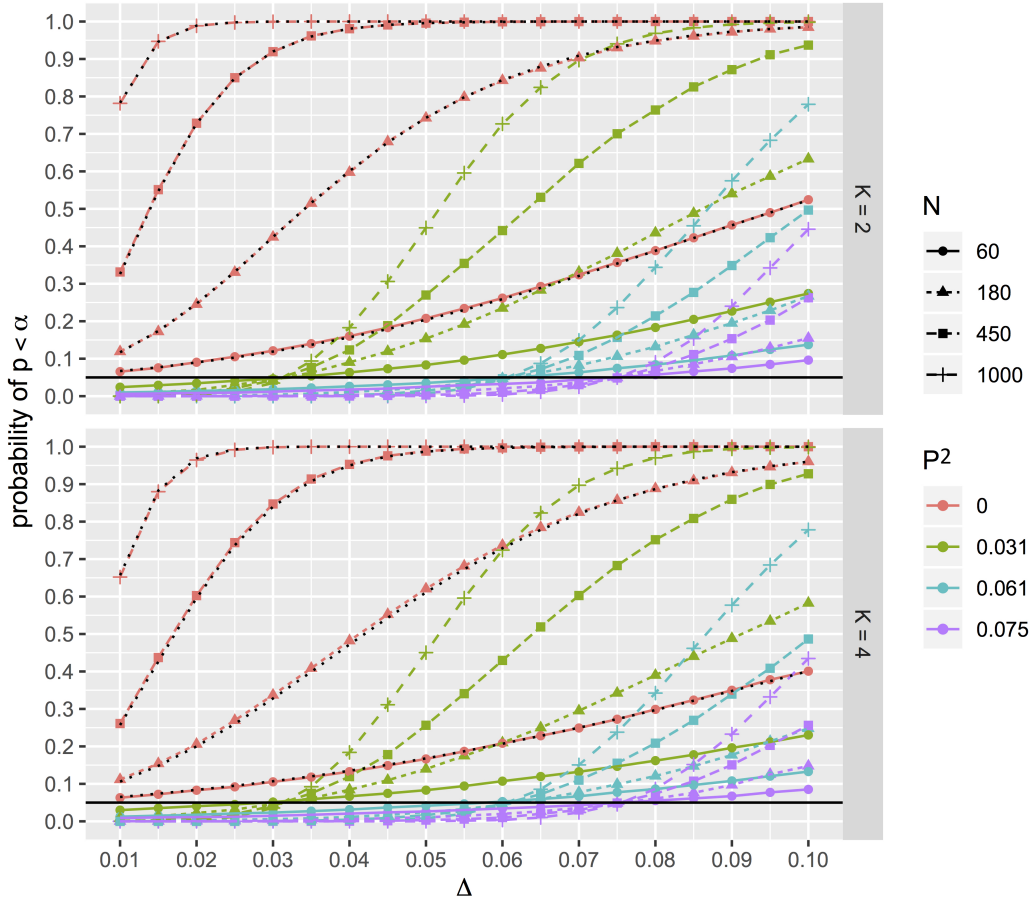

Figure 3. Simulation study, complete results. Upper panel shows results for $K = 2$; Lower panel shows results for $K = 4$. The solid horizontal black line indicates the desired type 1 error of $\alpha = 0.05$. The dotted black curves plot numbers calculated using equation (6) for estimating power. For each of thirty-two configurations, we simulated 50,000 unique datasets and calculated a non-inferiority p-value with each of 19 different values of Δ (ranging from 0.01 to 0.10).

for these data is either $P^2 = 0.031$, $P^2 = 0.061$, or $P^2 = 0.075$. Parameters for the simulation study were chosen so that we would consider a wide range of values for the sample size (representative of those sample sizes commonly used in the psychology literature; see Kühberger et al. (2014), Fraley and Vazire (2014), and Marszalek et al. (2011)) and so as to obtain three unique values for P^2 approximately evenly spaced between 0 and 0.10.

We also simulated data from eight additional scenarios where $P^2 = 0$. This will allow us to confirm that the proposed function (equation (6)) for estimating power is accurate. These additional scenarios were based on the following:

- one of four sample sizes: $N = 60$, $N = 180$, $N = 540$, or, $N = 1,000$;
- one of two designs with $K = 2$, or $K = 4$ binary covariates, with $\beta = (0.0, 0.0, 0.0)$ or $\beta = (0.0, 0.0, 0.0, 0.0, 0.0)$; and $\sigma^2 = 1.0$.

For each of the thirty-two configurations, we simulated 50,000 unique datasets and calculated a non-inferiority p-value with each of 19 different values of Δ (ranging from 0.01 to 0.10). We then calculated the proportion of these p-values less than $\alpha = 0.05$. We specifically chose to conduct 50,000 simulation runs so as to keep computing time within a reasonable limit while also reducing the amount of Monte Carlo standard error to a negligible amount (for looking at type 1 error with $\alpha = 0.05$, Monte Carlo SE will be approximately $0.001 \approx \sqrt{0.05(1 - 0.05)/50,000}$); see Morris et al. (2019).

Figure 2 plots the results with a restricted vertical axis to better show the type 1 error rates. Figure 3 plots the results against the unrestricted vertical axis. Both plots also show dotted black curves which correspond to the numbers obtained using equation (6) for calculating power.

We see that when the equivalence bound Δ equals the true effect size (i.e., 0.031, 0.061, or 0.075), the type 1 error rate is exactly 0.05, as it should be, for all N. This situation represents the boundary of the null hypothesis, i.e. $H_0 : \Delta \leq P^2$. As the equivalence bound increases beyond the true effect size (i.e., $\Delta > P^2$), the alternative hypothesis is then true and it becomes possible to correctly conclude equivalence.

As expected, the power of the test increases with larger values of Δ, larger values

of N, and smaller values of K. Also, in order for the test to have substantial power, the P^2 must be substantially smaller than Δ. The agreement between the red curves ($P^2 = 0$) and the dotted black curves suggests that the analytic function presented in equation (6) provides a fairly accurate approximation of the statistical power.

2.3. Simulation study 2

We conducted a second simulation study to compare the operating characteristics of testing with the JZS-BF relative to testing with the frequentist CET approach. (Note that the frequentist and Bayesian testing schemes we consider are but two of many options available to researchers. For example, one could consider a Bayesian approach that uses an interval-based null; see Kruschke and Liddell (2018).)

In this simulation study, frequentist conclusions were based on setting Δ equal to either 0.01, or 0.05, or 0.10; and with α=0.05. JZS-BF conclusions were based on an evidence threshold of either 3, 6, or 10. A threshold of 3 can be considered "moderate evidence," a threshold of 6 can be considered "strong evidence," and a threshold of 10 can be considered "very strong evidence" (Jeffreys, 1961; Wagenmakers et al., 2011). Note that for the simulation study here we examine only the "fixed-n design" for BF testing; see Schönbrodt and Wagenmakers (2016) for details. Also note that, as in Section 2.1, all priors required for calculating the BF were set by simply selecting the default settings of the `linearReg.R2stat()` function (with rscale = "medium"), whereby a noninformative Jeffreys prior is placed on the variance of the normal population, while a scaled Cauchy prior is placed on the standardized effect size; see Morey et al. (2015).

We simulated datasets for 64 unique scenarios. We considered the following parameters:

- one of sixteen sample sizes: $N = 20$, $N = 30$, $N = 42$, $N = 60$, $N = 88$, $N = 126$, $N = 182$, $N = 264$, $N = 380$, $N = 550$, $N = 794$, $N = 1,148$, $N = 1,658$, $N = 2,396$, $N = 3,460$, or $N = 5,000$;
- one of two designs with $K = 4$ binary covariates (with an orthogonal, balanced

design), with either $\beta = (0.0, 0.2, 0.2, -0.1, -0.2)$ or $\beta = (0.0, 0.0, 0.0, 0.0, 0.0)$;

- one of three variances: $\sigma^2 = 9.0, \sigma^2 = 1.0$, or $\sigma^2 = 0.5$.

Note that for the $\beta = (0.0, 0.0, 0.0, 0.0, 0.0)$ design, we only consider one value for $\sigma^2 = 1.0$. Depending on the particular design and σ^2, the true coefficient of determination for these data is either $P^2 = 0.000$, $P^2 = 0.004$, $P^2 = 0.031$, or $P^2 = 0.061$.

For each simulated dataset, we obtained frequentist p-values, JZS-BFs and declared the result to be positive, negative or inconclusive accordingly. Results are presented in Figures 4, 5 and 6 and are based on 5,000 distinct simulated datasets per scenario. We are also interested in how often the two approaches will reach the same overall conclusion: *averaging over all 64 scenarios, how often on average will the Bayesian and frequentist approaches reach the same conclusion given the same data?* Table 1 displays the the average rate of agreement between the Bayesian and frequentist methods.

Three observations merit comment:

- With an evidence threshold of 3 or of 6, the JZS-BF requires substantially less data to reach a negative conclusion than the frequentist scheme in most cases. However, with an evidence threshold of 10 and $\Delta = 0.10$, both methods are approximately equally likely to deliver a negative conclusion. Note that, the probability of reaching a negative result with CET will never exceed 0.95 since the NHST is performed first and will reach a positive result with probability $1 - \alpha$; see dashed orange lines in Figures 4, 5, and 6 - panels 1, 2, and 3.

- While the JZS-BF requires less data to reach a conclusive result when the sample size is small (see how the solid black curve drops more rapidly than the dashed grey line), there are scenarios in which larger sample sizes will surprisingly reduce the likelihood of the BF obtaining a conclusive result (see how the solid black curve drops abruptly then rises slightly as n increases for $P^2 = 0.004$, and 0.031; see for example, Figure 6 - panels 4 and 7).

- The JZS-BF is always less likely to deliver a positive conclusion (see how the dashed blue curve is always higher than the solid blue curve). In scenarios like

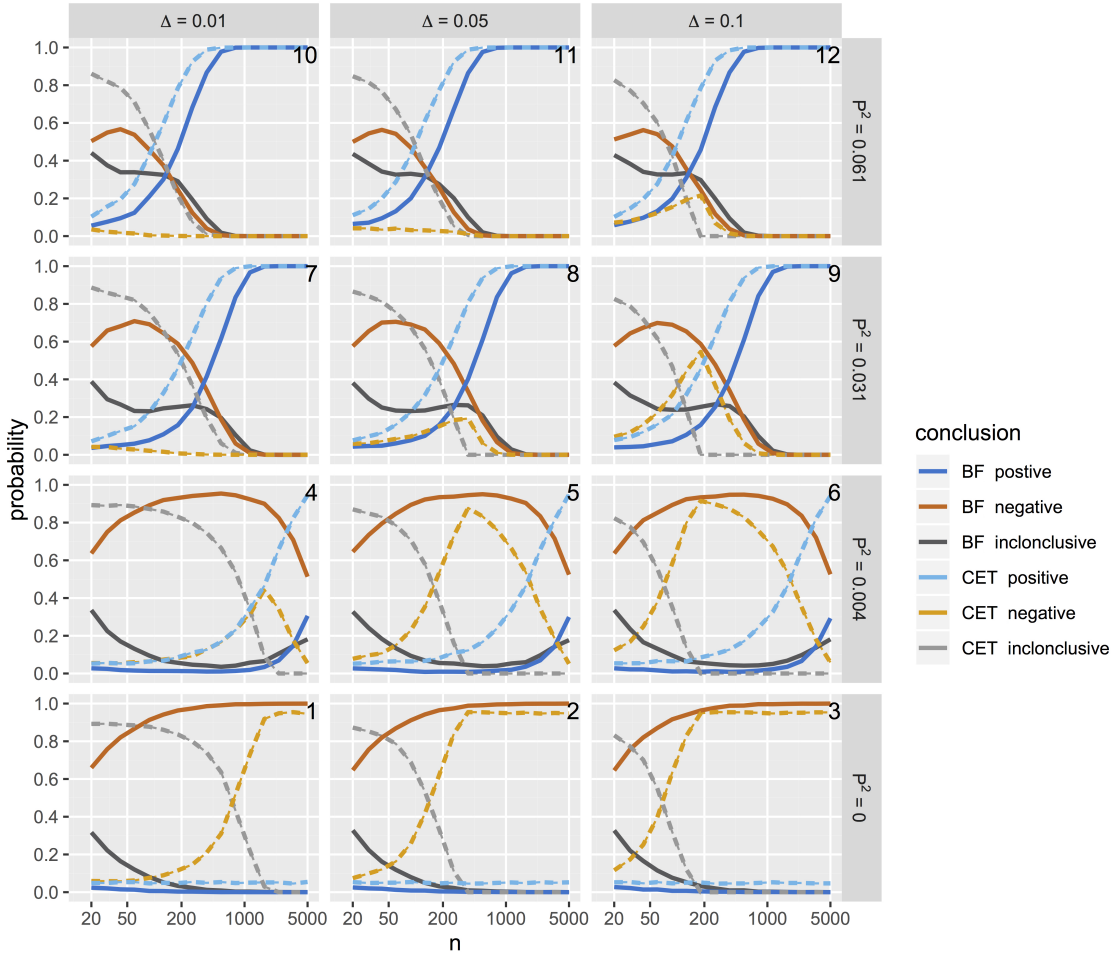

Figure 4. Simulation study 2, complete results for BF threshold of 3. The probability of obtaining each conclusion by Bayesian testing scheme (JZS-BF with fixed sample size design, BF threshold of 3:1) and CET ($\alpha = 0.05$). Each panel displays the results of simulations with for different values of Δ and P^2. Note that all solid lines and the dashed blue line do not change for different values of Δ.

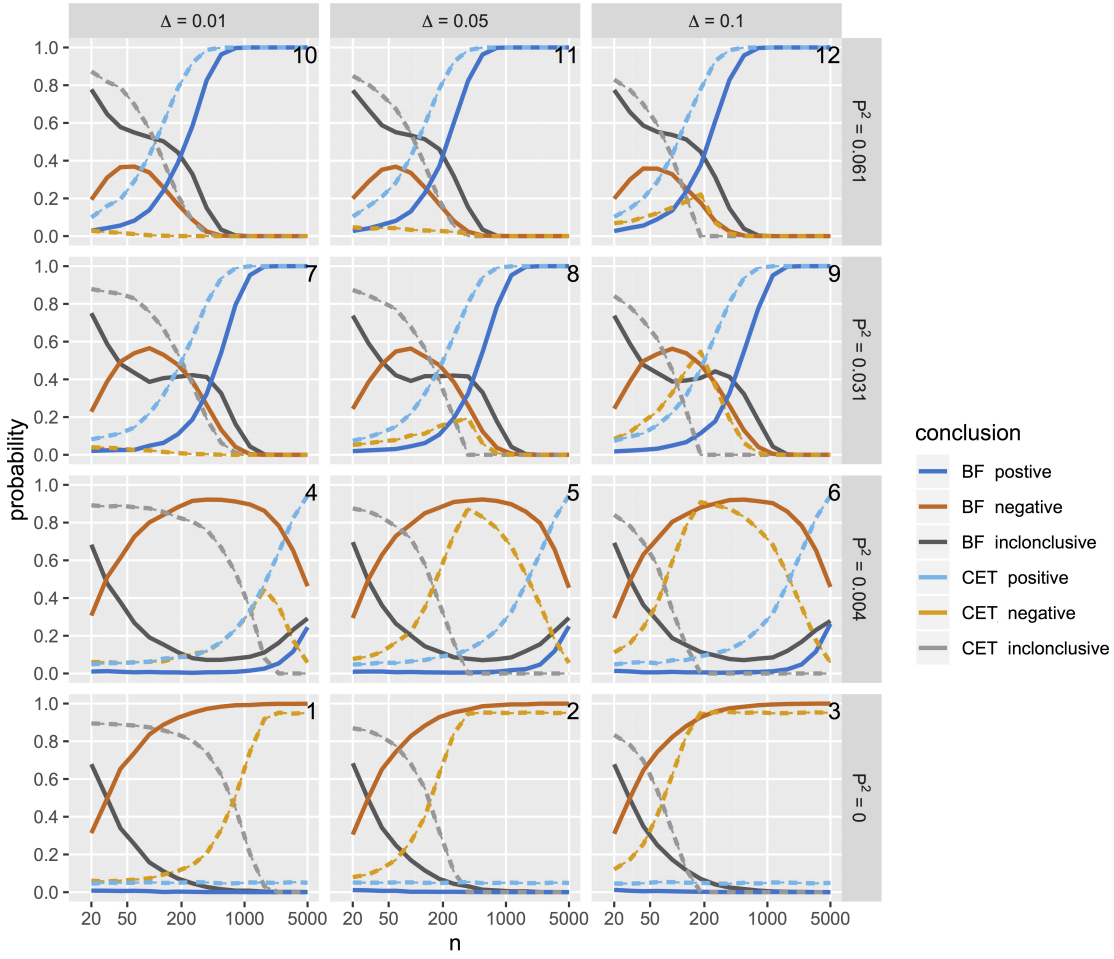

Figure 5. Simulation study 2, complete results for BF threshold of 6. The probability of obtaining each conclusion by Bayesian testing scheme (JZS-BF with fixed sample size design, BF threshold of 6:1) and CET ($\alpha = 0.05$). Each panel displays the results of simulations with for different values of Δ and P^2. Note that all solid lines and the dashed blue line do not change for different values of Δ.

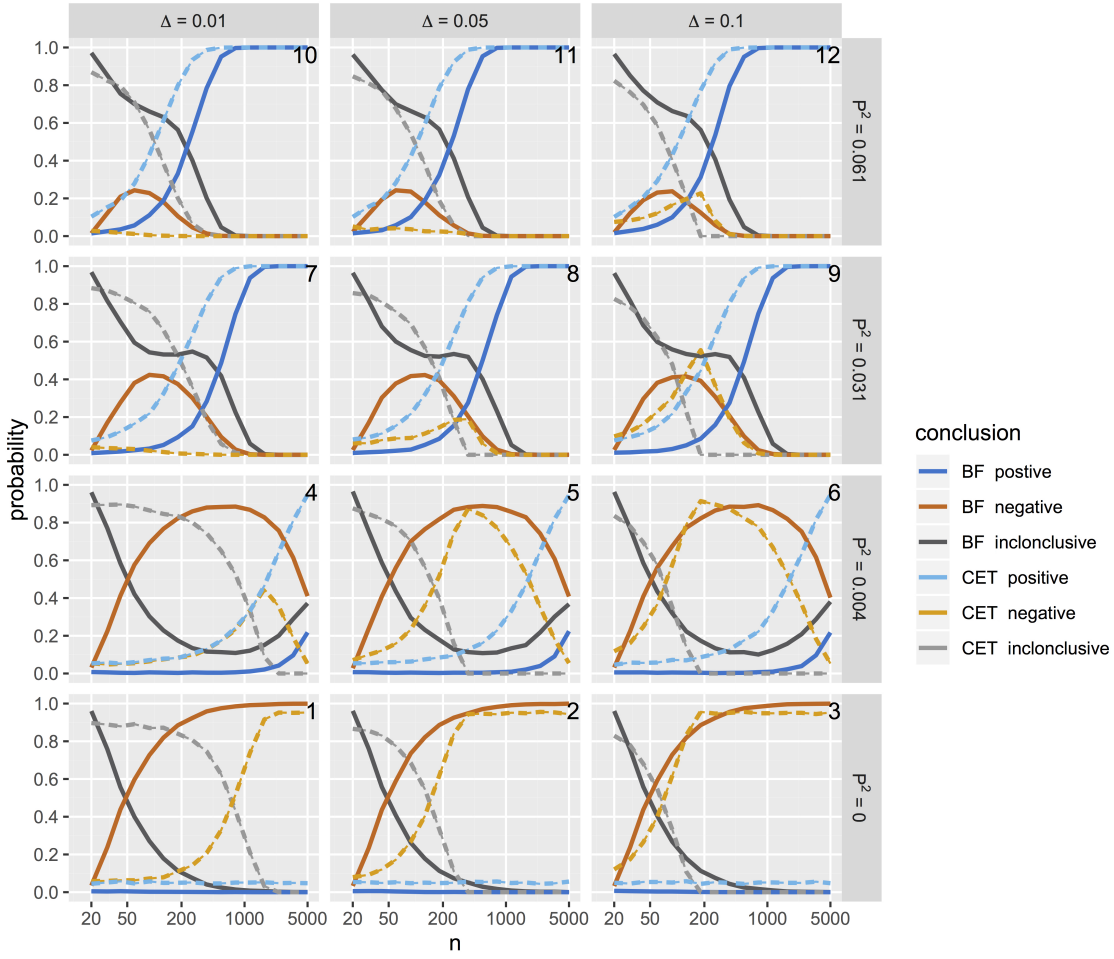

Figure 6. Simulation study 2, complete results for BF threshold of 10. The probability of obtaining each conclusion by Bayesian testing scheme (JZS-BF with fixed sample size design, BF threshold of 10:1) and CET ($\alpha = 0.05$). Each panel displays the results of simulations with for different values of Δ and P^2. Note that all solid lines and the dashed blue line do not change for different values of Δ.

the ones we considered, the JZS-BF may require larger sample sizes for reaching a positive conclusion and thus may be considered "less powerful" in a traditional frequentist sense.

Based on our comparison of BFs and frequentist tests, we can confirm that, given the same data, both approaches will often provide one with the same overall conclusion. The level of agreement however is highly sensitive to the choice of Δ and the choice of the BF evidence threshold, see Table 1. The highest level of agreement, recorded at 0.80, is when $\Delta = 0.10$ and the BF evidence threshold is equal to 10. In contrast, when $\Delta = 0.01$ and the BF evidence threshold is 3, the two approaches will only deliver the same conclusion half of the time. Table 2 shows that the two approaches rarely arrive at entirely contradictory conclusions. In less than 6% of cases, did we observe one approach arrive at a positive conclusion while the other approach arrived at a negative conclusion when faced with the same exact data.

The results of this second simulation study suggest that, depending on how they are configured, the JZS-BF and CET may operate very similarly. Think of JZS-BF and CET as two pragmatically similar, yet philosophically different, tools for making "trichotomous significance-testing decisions." This simulation study result is reassuring since it suggests that the conclusions obtained from frequentist and Bayesian testing will rarely lead to substantial disagreements.

	BF thres. = 3	BF thres. = 6	BF thres. = 10
$\Delta = 0.10$	0.719	0.767	0.800
$\Delta = 0.05$	0.628	0.683	0.735
$\Delta = 0.01$	0.485	0.538	0.594

Table 1. Averaging over all 96 scenarios and over all 5,000 Monte Carlo simulations per scenario, how often on average did the Bayesian and frequentist approaches reach the same conclusion? Numbers in the above table represent the average proportion of simulated datasets (averaged over $64 \times 5,000 = 320,000$ unique datasets) for which the Bayesian and frequentist methods arrive at the same conclusion.

	BF thres. = 3	BF thres. = 6	BF thres. = 10
$\Delta = 0.10$	0.055	0.042	0.034
$\Delta = 0.05$	0.055	0.042	0.034
$\Delta = 0.01$	0.056	0.042	0.035

Table 2. Averaging over all 96 scenarios and over all 5,000 Monte Carlo simulations per scenario, how often on average did the Bayesian and frequentist approaches strongly disagree in their conclusion? Numbers in the above table represent the average proportion of simulated datasets (averaged over $64 \times 5,000 = 320,000$ unique datasets) for which the Bayesian and frequentist methods arrived at completely opposite (one positive and one negative) conclusions.

3. Practical Example: Evidence for the absence of a Hawthorne effect

McCambridge et al. (2019) tested the hypothesis that participants who know that the behavioral focus of a study is alcohol related will modify their consumption of alcohol while under study. The phenomenon of subjects modifying their behaviour simply because they are being observed is commonly known as the Hawthorne effect (Stand, 2000).

The researchers conducted a three-arm individually randomized trial online among students in four New Zealand universities. The three groups were: group A (control), who were told they were completing a lifestyle survey; group B, who were told the focus of the survey was alcohol consumption; and group C, who additionally answered 20 questions on their alcohol use and its consequences before answering the same lifestyle questions as Groups A and B. The prespecified primary outcome was a subject's self-reported volume of alcohol consumption in the previous 4 weeks (units = number of standard drinks). This measure was recorded at baseline and after one month at follow-up. See Table 3 for a summary of the data from McCambridge at el. (2019).

The data were analyzed by McCambridge et al. (2019) using a linear regression model with repeated measures fit by generalized estimating equations (GEE) and an "independence" correlation structure. For a NHST of the overall experimental group effect, the researchers obtained a p-value of 0.66. Based on this result, McCambridge et al. (2019) conclude that "the groups were not found to change differently over time."

We note that this linear regression model fit by GEE is just one of many potential models one could use to analyze this data; see Yang and Tsiatis (2001). Three (among

		baseline	followup	difference
A	N	1795	1483	1483
	mean	24.60	18.39	-5.13
	sd	31.80	23.32	24.56
B	N	1852	1532	1532
	mean	23.83	17.48	-5.64
	sd	31.79	23.81	21.77
C	N	1825	1565	1565
	mean	23.03	17.45	-4.79
	sd	30.65	23.21	25.17
Total	N	5472	4582	4580
	mean	23.82	17.77	-5.19
	sd	31.42	23.44	23.88

Table 3. Summary of the data from McCambridge at el. (2019). The table summarizes the prespecified primary outcome, a subject's self-reported volume of alcohol consumption in the previous 4 weeks (units = number of standard drinks). This measure was recorded at baseline and after one month at follow-up in each of the three experimental groups.

many) other reasonable alternative approaches include (1) a linear model using only the follow-up responses (without adjustment for the baseline measurement); (2) a linear model using the follow-up responses as outcome with a covariate adjustment for the baseline measurement; and (3) a linear model using the difference between follow-up and baseline responses as outcome. These three approaches yield p-values of 0.45, 0.56, and 0.61, respectively. None of these p-values suggest rejecting the null hypothesis. In order to show evidence "in favour of the null," we turn to our proposed non-inferiority test.

We fit the data ($N = 4,580$) with a linear regression model using the difference between follow-up and baseline responses as the outcome, and the group membership as a categorical covariate, $K = 2$. We then consider the non-inferiority test for the coefficient of determination parameter (see Section 2), with $\Delta = 0.01$. This test asks the following question: does the overall experimental group effect account for less than 1% of the variability explained in the outcome?

The choice of $\Delta = 0.01$ represents our belief that any Hawthorne effect explaining less than 1% of the variability in the data would be considered negligible. For reference, Cohen (1988) describes a $R^2 = 0.0196$ as "a modest enough amount, just barely escaping triviality"; and more recently, Fritz et al. (2012) consider associations explaining "1% of the variability" as "trivial." It is up to researchers to provide a justi-

fication of the equivalence bound before they collect the data. Researchers can specify the non-inferiority margin based on theoretical predictions (for example derived from a computational model), based on a cost-benefit analysis, or based on discussions among experts who decide which effects are too small to be considered meaningful.

We obtain a $R^2 = 0.000216$ and can calculate the F-statistic with equation (4):

$$
\begin{aligned}
F = & \frac{R^2/K}{(1-R^2)/(N-K-1)} && (10)\\
= & \frac{0.000216/2}{(1-0.000216)/(4580-2-1)} && (11)\\
= & \frac{0.000108}{0.000218} && (12)\\
= & 0.49 && (13)
\end{aligned}
$$

To obtain a p-value for the non-inferiority test, we use equation (5):

$$
\begin{aligned}
p-\text{value} = & p_f\left(F; K, N-K-1, \frac{N\Delta}{(1-\Delta)}\right) && (14)\\
= & p_f\left(0.49; 2, 4580-2-1, \frac{4580\cdot 0.01}{(1-0.01)}\right) && (15)\\
= & 1.13\times 10^{-9} && (16)
\end{aligned}
$$

This result, p-value $= 1.13\times 10^{-9}$, suggests that we can confidently reject the null hypothesis that $P^2 > 0.01$. We therefore conclude that the data are most compatible with no important effect. For comparison, the Bayesian testing scheme we considered in Section 2.1 obtains a Bayes Factor of $B_{10} = 0.00284 = 1/352$. The `R` code for these calculations is presented in the Appendix.

4. A non-inferiority test for the ANOVA η^2 parameter

Despite being entirely equivalent to linear regression (Gelman et al., 2005), the fixed effects (or "between subjects") analysis of variance (ANOVA) continues to be the most common statistical procedure to test the equality of multiple independent population means in many fields (Plonsky and Oswald, 2017). The non-inferiority test considered earlier in the linear regression context will now be described in an ANOVA context for evaluating the equivalence of multiple independent groups. We must emphasize that the two versions are essentially the same test described with different names. Note that all tests developed and discussed in this paper are only for *between*-subject ANOVA designs and cannot be applied to *within*-subject designs. Extensions to within and mixed designs is a fruitful effort for future research.

Equivalence/non-inferiority tests for comparing group means in an ANOVA have been proposed before. For example, Rusticus and Lovato (2011) list several examples of studies that used ANOVA to compare multiple groups in which non-significant findings are incorrectly used to conclude that groups are comparable. The authors emphasize the problem ("a statistically non-significant finding only indicates that there is not enough evidence to support that two (or more) groups are statistically different") and offer an equivalence testing solution based on CIs. Unfortunately, conclusions of equivalence are based only on CIs which the authors warn may be "too wide" (Rusticus and Lovato, 2011).

In another proposal, Wellek (2010) considered simultaneous equivalence testing for several parameters to test group means. However, this strategy may not necessarily be more efficient than the rather inefficient strategy of multiple pairwise comparisons; see the conclusions of Pallmann and Jaki (2017). Koh and Cribbie (2013) (see also Cribbie et al. (2009)) consider two different omnibus tests. These are presented as non-inferiority tests for φ^2, a parameter closely related to the population signal-to-noise parameter, $s2n$; (note that $s2n = \varphi^2/N$, where N is the total sample size). Unfortunately, the use of these tests is limited by the fact that the population parameters φ^2 and $s2n$ are not commonly used in analyses since their units of measurement

are rather arbitrary.

In this section, we consider a non-inferiority test for the population effect-size parameter, η^2, a standardized effect size that is commonly used in the social sciences (Kelley et al., 2007). The parameter η^2 represents the proportion of total variance in the population that can be accounted for by knowing the group level. The use of commonly used standardized effect sizes is recommended in order to facilitate future meta-analysis and the interpretation of results (Lakens, 2013). Note that η^2 is analogous to the P^2 parameter considered earlier in the linear regression context in Section 2. Also note that the non-inferiority test we propose is entirely equivalent to the test for φ^2 proposed by Koh and Cribbie (2013). It is simply a re-formulation of the test in terms of the η^2 parameter.

Before going forward, let us define some basic notation. All technical details are presented in the Appendix. Let Y represent the continuous (normally distributed) outcome variable, and X represent a fixed categorical variable (i.e., group membership). Let N be the total number of observations in the observed data, J be the number of groups (i.e., factor levels in X), and n_j be the number of observations in the jth group, for j in 1,..., J. We will consider two separate cases, one in which the variance within each group is equal, and one in which variance is heterogeneous.

Typically, one will conduct a standard F-test to determine whether one can reject the null hypothesis that η^2 is equal to zero ($H_0 : \eta^2 = 0$). The p-value is calculated as:

$$p - \text{value} = 1 - p_f(F; J-1, N-J, 0), \tag{17}$$

where, as in Section 2, $p_f(\cdot \quad ; df_1, df_2, ncp)$ is the cdf of the non-central F-distribution with df_1 and df_2 degrees of freedom, and non-centrality parameter, ncp (note that $ncp = 0$ corresponds to the *central* F-distribution); and where:

$$F = \frac{\sum_{j=1}^{J} n_j(\bar{y}_j - \bar{y})^2/(J-1)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2/(N-J)}. \tag{18}$$

One can calculate the above p-value using `R` with the following code:

```
pval = pf(Fstat,df1=J-1,df2=N-J,lower.tail=FALSE).
```

A non-inferiority test for η^2 asks a different question: *can we reject the hypothesis that the total amount of variance in Y attributable to group membership is greater than Δ?* Formally, the hypotheses for the non-inferiority test are written as:

$H_0 : 1 > \eta^2 \geq \Delta,$
$H_1 : 0 < \eta^2 \leq \Delta.$

If we reject H_0, we reject the hypothesis that there are meaningful differences between the group means (μ_j, $j = 1, ..., J$), in favour of the hypothesis that the group means are considered practically equivalent. The p-value for this test is obtained by inverting the one-sided CI for η^2 (see Appendix for details) and can be calculated as:

$$p - \text{value} = p_f\left(F; J-1, N-J, \frac{N\Delta}{(1-\Delta)}\right). \tag{19}$$

Note that one can calculate the above p-value using `R` with the following code:

```
pval = pf(Fstat,df1=J-1,df2=N-J,ncp=N*Delta/(1-Delta),lower.tail=TRUE).
```

Under the assumption that the true value of $\eta^2 = 0$, for given values of N, J, and Δ, a simple analytic formula provides an estimate for the non-inferiority test's statistical power:

$$power = Pr(\text{reject } H_0 | \eta^2 = 0) = p_f(F^*; J-1, N-J, 0), \tag{20}$$

where F^* is equal to the $(1-\alpha)\%$ critical value of a non-central F-distribution with $df_1 = J-1$ and $df_2 = N-J$ degrees of freedom, and non-centrality parameter $ncp = (N\Delta)/(1-\Delta)$.

Note that one can calculate the above power estimate in `R` with the following code:

```
Fstatstar = qf(alpha, df1 = J-1, df2 = N-J, ncp = (N*Delta)/(1-Delta), lower.tail = TRUE)

power = pf(Fstatstar, df1 = J-1, df2 = N-J, lower.tail = TRUE).
```

The non-inferiority test for η^2 makes the following three important assumptions about the data:

- The outcome data are independent and normally distributed.
- The proportions of observations for each group (i.e., n_j/N, for $j = 1, ...J$) that are in the observed data are equal to the proportions that are in the total population of interest.
- The variance within each group is equal (homogeneous variance).

4.1. A non-inferiority test for ANOVA with heterogeneous variance

With regards to the third assumption above, we can modify the above non-inferiority test in order to allow for the possibility that the variance is unequal across groups (heterogeneous variance). Welch's F-test statistic is calculated as (see Appendix for details; see also Delacre et al. (2018)):

$$F' = \frac{\sum_j^J \hat{w}_j(y_j - \bar{y}')^2/(J-1)}{1 + \frac{2(J-2)}{J^2-1}\sum_{j=1}^J((n_j-1)^{-1})(1-\frac{\hat{w}_j}{\hat{W}})}, \tag{21}$$

where $\hat{w}_j = n_j/s_j^2$, with $s_j^2 = \sum_{i=1}^{n_j}((y_{ij} - \bar{y}_j)^2)/(n_j - 1)$, for $j = 1, ..., J$; and where $\hat{W} = \sum_{j=1}^J \hat{w}_j$, and $\bar{y}' = \sum_{j=1}^J(\hat{w}_j\bar{y}_j)/\hat{W}$, for $j = 1, ..., J$.

Then, the p-value for a non-inferiority test ($H_0 : 1 > \eta^2 \geq \Delta$) in the case of heterogeneous variance is:

$$p - \text{value} = p_f(F^{'}; J-1, df^{'}, \frac{N\Delta}{(1-\Delta)}). \tag{22}$$

where:

$$df^{'} = \frac{J^2 - 1}{3\sum_{j=1}^{J}((n_j - 1)^{-1})(1 - \hat{w}_j/\hat{W})^2}. \tag{23}$$

The above p-value can be calculated using `R` with the following code:

```
aov1 <- oneway.test(y ~ x, var.equal = FALSE)
Fprime <- aov1$statistic
dfprime <-  aov1$parameter[2]
pval = pf(Fprime, J-1, df2 = dfprime, ncp = (Delta*N)/(1-Delta), lower.tail=TRUE)
```

For the heterogeneous case the population effect size parameter, η^2, is defined slightly differently than for the homogeneous case (see Appendix for details). Based on the simulation studies of Koh and Cribbie (2013), we can recommend that the non-inferiority test based on the Welch's $F^{'}$ statistic (i.e., the test with p-value calculated from equation (22)) is almost always preferable (with regards to the statistical power and type 1 error rate) to the test which requires an assumption of homogeneous variance (i.e., the test with p-value calculated from equation (19)). This agrees with similar recommendations for using Welch's t-test (e.g., Delacre et al. (2017); Ruxton (2006)). We also point interested readers to the related work of Jan and Shieh (2019).

Some might advocate for a two-step procedure: using the homoscedastic version as a default and only moving to the Welch version as needed based on a preliminary test for homogeneity of variance. However, problems with this kind of preliminary testing (e.g., first testing for equality of variances, then deciding upon which test to use) have been identified (e.g., Zimmerman (2004a,b); Campbell and Dean (2014)),

and as such, the use of the two-step procedure cannot be recommended.

4.2. The absence of a Hawthorne effect (ANOVA)

For the absence of a Hawthorne effect example we considered earlier in Section 3, note that we can easily analyze the data in an ANOVA framework (and obtain the identical result). The standard ANOVA output is summarized in Table 4.

	df	Sum Sq	Mean Sq	F value
Experimental Group	2	562.77	281.38	0.49
Residuals	4577	2609820.50	570.20	
				$\hat{\eta}^2 = 0.000216$

Table 4. ANOVA summary of the absence of a Hawthorne effect example data.

In this case, with $\Delta = 0.01$, a non-inferiority test can be conducted for η^2 ($H_0 : 1 > \eta^2 \geq \Delta$) and a p-value is calculated using equation (19) as follows:

$$p - \text{value} = p_f\left(F; J-1, N-J, \frac{N\Delta}{(1-\Delta)}\right) \tag{24}$$

$$= p_f\left(0.49; 3-1, 4580-3, \frac{4580 \cdot 0.01}{(1-0.01)}\right) \tag{25}$$

$$= 1.13 \times 10^{-9}. \tag{26}$$

5. Conclusion

In this paper we presented a statistical method for non-inferiority testing of standardized omnibus effects commonly used in linear regression and ANOVA. We also considered how frequentist non-inferiority testing, and equivalence testing more generally, offer an attractive alternative to Bayesian methods for "testing the null." We recommend that all researchers specify an appropriate non-inferiority margin and, at a minimum, plan to use the proposed non-inferiority tests in the event that a standard NHST fails to reject the null. In cases when the sample size is very large, the non-inferiority test can be useful to detect effects that are statistically significant but

not meaningful.

We wish to emphasize that the use of equivalence/non-inferiority tests should not rule out the complementary use of confidence intervals. Indeed, confidence intervals can be extremely useful for highlighting the stability (or lack of stability) of a given estimator, whether that be the R^2, $\hat{\eta}^2$ or any other statistic. Perhaps one advantage of equivalence/non-inferiority testing over confidence intervals may be that testing can improve the interpretation of null results (Parkhurst, 2001; Hauck and Anderson, 1986). By clearly distinguishing between what is a "negative" versus an "inconclusive" result, equivalence testing serves to simplify the long "series of searching questions" necessary to evaluate a "failed outcome" (Pocock and Stone, 2016). In our opinion, the best interpretation of data will be when using both tools together and our proposal simply serves to "extend the arsenal of confirmatory methods rooted in the frequentist paradigm of inference" (Wellek, 2017).

Note that our current non-inferiority test for P^2 in a standard multivariable linear regression is limited to comparing the "full model" to the "null model." As such, the test is not suitable for comparing two nested models. For example, we cannot use the test to compare a "smaller model" with only the baseline measure as a covariate, with a "larger model" that includes both baseline measure *and* group membership as covariates. Equivalence testing for comparing two nested models will be addressed in future work in which we will consider a non-inferiority test for the increase in R^2 between a smaller model and a larger model. Related work includes that of Algina et al. (2007, 2008). We also wish to further investigate non-inferiority testing for ANOVA with *within*-subject designs, following the work of Rose et al. (2018). It would also be interesting (and worthwhile) to develop non-inferiority tests to tackle the R^2 calculated for generalized linear mixed-effects models (Nakagawa and Schielzeth, 2013).

There is a great risk of bias in the scientific literature if researchers only rely on statistical tools that can reject null hypotheses, but do not have access to statistical tools that allow them to reject the presence of meaningful effects. Most recently, Amrhein et al. (2019) express great concern with the the practice of statistically non-significant results being "interpreted as indicating 'no difference' or 'no effect' " (Am-

rhein et al., 2019); see also Altman and Bland (1995). Equivalence tests provide one approach to improve current research practices by allowing researchers to falsify their predictions concerning the presence of an effect.

By specifying equivalence bounds, researchers can design studies that yield informative answers both when the alternative hypothesis is true, and when the null hypothesis is true. Using equivalence tests to reject the presence of meaningful effects makes it possible to conclude predictions are falsified, and thus might be a way to mitigate problems that are caused by publication bias. However, equivalence tests can also be abused. Researchers might be tempted to specify the equivalence bounds after looking at the data such that the equivalence test is guaranteed to be statistically significant. Ideally, equivalence bounds are pre-specified and documented in a preregistration document that is made available when a manuscript is submitted for publication to avoid flexibility in the data analysis. Equivalence bounds should always be justified and independent of the observed data. Weak justifications weaken the statistical inference. Thinking about what would falsify your prediction is a crucial step when designing a study, and specifying a smallest effect size of interest and performing an equivalence test provides one way to answer that question.

6. Appendix

6.1. Linear Regression: further details and R-code.

The R^2 statistic estimates the parameter P^2 from the observed data:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}}, \tag{27}$$

where $SS_{RES} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$, and $SS_{TOT} = \sum_{i=1}^{N} (y_i - \overline{y})^2$; with $\hat{y} = X^T (X'X)^{-1} X'Y$, and $\bar{y} = \sum_{i=1}^{N} y_i/N$.

The R-code for analysis of the McCambridge et al. (2019) data is:

```
Xmatrix <- model.matrix(totaldrinking.diff  ~ group, data= side_data)
lmmodel <- lm(totaldrinking.diff  ~ group , data= side_data)

R2 <- summary(lmmodel)$r.squared
Fstat <- summary(lmmodel)$fstatistic[1]
K <- dim(Xmatrix)[2] - 1
N <- dim(Xmatrix)[1]
Delta <- 0.01

pf(Fstat,df1=K,df2=N-K-1,ncp=(N*Delta)/(1-Delta),lower.tail=TRUE)

linearReg.R2stat(N=N, p=K, R2= R2, simple=TRUE)
```

The code below replicates the results published in McCambridge et al. (2019). Note that there appears to be a typo in the published table whereby the p-values 0.89 and 0.86 are switched.

```
Hdata$group<-relevel(Hdata$group,"A")

mod0 <- geeglm(totaldrinking ~ + group+t,
id= participant_ID, corstr="independence", data= Hdata, x=TRUE)
mod1 <- geeglm(totaldrinking ~ group*t + group+t,
id= participant_ID, corstr="independence", data= Hdata)
(anova(mod1,mod0))
summary(mod1)$coefficients

Hdata$group<-relevel(Hdata$group,"C")
mod1a <- geeglm(totaldrinking ~ group*t + group+t,
```

```
id= participant_ID, corstr="independence", data= Hdata)
summary(mod1a)
```

6.2. ANOVA with homogeneous variance: further details.

The true population group mean for group j is denoted μ_j, for j in 1,..., J; and we denote the group effects as $\tau_j = \mu_j - \mu$, where μ is the overall weighted population mean, $\mu = (\sum_{j=1}^{J} \mu_j n_j)/N$. These parameters are estimated from the observed data by the corresponding sample group means: $\hat{\mu}_j = \bar{y}_j = (\sum_{i=1}^{n_j} y_i)/n_j$, for j in 1,...,J; and the overall sample mean: $\hat{\mu} = \bar{y} = (\sum_{j=1}^{J} \bar{y}_j n_j)/N$.

We operate under the assumption that the data is normally distributed such that:

$$Y_{i,j} \sim \ Normal(\mu_j, \sigma_w^2), \ \forall \ j = 1, ..., J, \quad \forall i = 1, ..., n_j, \tag{28}$$

where σ_w^2 denotes the variance within groups. We also define the variance between groups as $\sigma_b^2 = \sum_{j=1}^{J} n_j(\mu_j - \mu)^2/N$. Finally, the total population variance is defined as $\sigma_t^2 = \sigma_b^2 + \sigma_w^2$. The corresponding sums of squares are estimated from the data: $SS_b = \sum_{j=1}^{J} n_j(\bar{y}_j - \bar{y})^2$; $SS_w = \sum_{j=1}^{J} \sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$; and $SS_t = SS_b + SS_w$.

Recall that the ANOVA F-test statistic is calculated as:

$$F = \frac{SSb/df_b}{SS_w/df_w} = \frac{\sum_{j=1}^{J} n_j(\bar{y}_j - \bar{y})^2/(J-1)}{\sum_{j=1}^{J} \sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2/(N-J)}, \tag{29}$$

where $df_b = J - 1$, and $df_w = N - J$. The F statistic follows an F distribution with degrees of freedom df_b for the numerator, and df_w degrees of freedom for the denominator.

The population effect size, $\eta^2 \in [0, 1]$, is a parameter that represents the amount of variance in the outcome variable, Y, that is explained by the group membership,

(i.e., knowing the level of the factor X), and is defined as:

$$\eta^2 = \frac{\sigma_b^2}{\sigma_t^2} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} = 1 - \frac{\sigma_w^2}{\sigma_t^2} \tag{30}$$

We can estimate the population parameter η^2 from the observed data using the sample statistic, $\hat{\eta}^2$, as follows: $\hat{\eta}^2 = SS_b/SS_t$. It is well known that $\hat{\eta}^2$ is a biased estimate for η^2. However, alternative estimates (including $\hat{\epsilon}^2 = (SS_b - dfb \cdot MS_w)/SS_t$, and $\hat{\omega}^2 = (SS_b - df_b \cdot MS_w)/(SS_t + MS_w)$) are also biased; see Okada (2013) for more details (note that there is a typo in eq. 5 of Okada (2013)).

The population effect size parameter η^2 is closely related to the signal-to-noise ratio parameter, $s2n = \sigma_b^2/\sigma_w^2$, and to the non-centrality parameter, $\Lambda = \sum_{j=1}^{J} n_j \tau_j^2/\sigma_w^2 = N\sigma_b^2/\sigma_w^2$. Consider the following equality:

$$\eta^2 = \frac{s2n}{1 - s2n} = \frac{\Lambda}{\Lambda + N}. \tag{31}$$

The non-centrality parameter, Λ, is estimated from the data as: $\hat{\Lambda} = (n - 1)SS_b/SS_w$, and we can easily calculate a one-sided $(1-\alpha)\%$ confidence interval (CI), $[0, \Lambda_U]$, by "pivoting" the cumulative distribution function (cdf); see Kelley et al. (2007) Section 2.2 and references therein. This requires solving (numerically) the following equation for Λ_U:

$$p_f(F; df_1 = df_b, df_2 = df_w, ncp = \Lambda_U) = \alpha, \tag{32}$$

where $p_f(\cdot \quad ; df_1, df_2, ncp)$ is the cdf of the non-central F-distribution with df_1 and df_2 degrees of freedom, and non-centrality parameter, ncp. The values for F, df_b, df_w, are calculated from the data as defined above. The solution, Λ_U, will be the upper

confidence bound of Λ, such that: $Pr(\Lambda < \Lambda_U) = \alpha$.

As detailed in Kelley et al. (2007) (note that there is a typo in eq. 55 of Kelley et al. (2007): Λ_L in the numerator should be Λ_U) one can convert the bounds of the CI for Λ into bounds for a CI for η^2. The upper limit of a one-sided CI for η^2 is: $\eta^2_U = \Lambda_U/(\Lambda_U + N)$. As such, we have that $Pr(\eta^2 \leq \frac{\Lambda_U}{\Lambda_U + N}) = 1 - \alpha$.

6.3. ANOVA with heterogeneous variance: further details.

As above, the true population group mean for group j is denoted μ_j, for j in 1,...,J. We now define:

$$Y_{i,j} \sim Normal(\mu_j, \sigma^2_{w,j}),\ \forall j = 1, ..., J, \quad \forall i = 1, ..., n_j, \tag{33}$$

and define $w_j = n_j/\sigma^2_{w,j}$, and $W = \sum_{j=1}^{J} w_j$, and finally $\bar{\mu}^{'} = \sum_{j=1}^{J}(w_j\mu_j)/W$.

Recall that a Welch F-test statistic is calculated as:

$$F^{'} = \frac{\sum_j^J \hat{w}_j(\bar{y}_j - \bar{y}^{'})^2/(J-1)}{1 + \frac{2(J-2)}{J^2-1}\sum_{j=1}^{J}((n_j-1)^{-1})(1-\hat{w}_j/\hat{W})^2}, \tag{34}$$

where $\hat{w}_j = n_j/s^2_j$, with $s^2_j = \sum_{i=1}^{n_j}((y_{ij} - \bar{y}_j)^2)/(n_j - 1)$, for $j = 1, ..., J$; and where $\hat{W} = \sum_{j=1}^{J} \hat{w}_j$, and $\bar{y}^{'} = \sum_{j=1}^{J}(\hat{w}_j\bar{y}_j)/\hat{W}$, for $j = 1, ..., J$.

Levy (1978) proposed an approximate non-null distribution for the $F^{'}$ statistic such that $F^{'}$ follows a non-central F-distribution with $df_1 = J - 1$ and $df_2 = df^{'}$ degrees of freedom, and non-centrality parameter, $\Lambda^{'} = \sum_{j=1}^{J} w_j(\mu_j - \bar{\mu}^{'})^2$; see also Jan and Shieh (2014). The degrees of freedom for this case are defined as: $df_1 = J - 1$, and:

$$df^{'} = \frac{J^2 - 1}{3\sum_{j=1}^{J}((n_j - 1)^{-1})(1 - \hat{w}_j/\hat{W})^2} \tag{35}$$

We will therefore define our population effect size parameter for the heterogeneous case as:

$$\eta^{2'} = \frac{\Lambda^{'}}{\Lambda^{'} + N}. \tag{36}$$

Note that in the case of homogeneous variance (i.e., when $\sigma^2_{w,j} = \sigma^2_{w,k}, \forall j, k$ in $1, ..., J$), we have $\Lambda^{'} = \Lambda$ and $\eta^{2'} = \eta^2$. The p-value for the non-inferiority test ($H_0 : \eta^{2'} > \Delta$) in the case of heterogeneous variance is:

$$p - \text{value} = p_f\left(F^{'}; J - 1, df^{'}, ncp = \frac{N\Delta}{(1 - \Delta)}\right). \tag{37}$$

References

Algina, J., Keselman, H., and Penfield, R. D. (2007). Confidence intervals for an effect size measure in multiple linear regression. *Educational and Psychological Measurement*, 67(2):207–218; https://doi.org/10.1177/0013164406292030.

Algina, J., Keselman, H. J., and Penfield, R. J. (2008). Note on a confidence interval for the squared semipartial correlation coefficient. *Educational and Psychological Measurement*, 68(5):734–741; https://doi.org/10.1177/0013164407313371.

Altman, D. G. and Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *The BMJ*, 311(7003):485; https://doi.org/10.1136/bmj.311.7003.485.

Amrhein, V., Greenland, S., and McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, (567):305–307; https://doi.og/10.1038/d41586–019–00857–9.

Barten, A. (1962). Note on unbiased estimation of the squared multiple correlation coefficient. *Statistica Neerlandica*, 16(2):151–164.

Campbell, H. and Dean, C. (2014). The consequences of proportional hazards based model selection. *Statistics in Medicine*, 33(6):1042–1056.

Campbell, H. and Gustafson, P. (2018a). Conditional equivalence testing: An alternative remedy for publication bias. *PLoS ONE*, 13(4):e0195145; https://doi.org/10.1371/journal.pone.0195145.

Campbell, H. and Gustafson, P. (2018b). What to make of non-inferiority and equivalence testing with a post-specified margin? *arXiv preprint arXiv:1807.03413*.

Consonni, G., Veronese, P., et al. (2008). Compatibility of prior specifications across linear models. *Statistical Science*, 23(3):332–353.

Cramer, J. S. (1987). Mean and variance of R2 in small and moderate samples. *Journal of Econometrics*, 35(2-3):253–266.

Cribbie, R. A., Arpin-Cribbie, C. A., and Gruman, J. A. (2009). Tests of equivalence

for one-way independent groups designs. *The Journal of Experimental Education*, 78(1):1–13; https://doi.org/10.1080/00220970903224552.

Delacre, M., Lakens, D., and Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology*, 30(1).

Delacre, M., Lakens, D., Mora, Y., and Leys, C. (2018). Taking parametric assumptions seriously arguments for the use of Welch's F-test instead of the classical F-test in one-way ANOVA; https://doi.org/10.31234/osf.io/wnezg.

Dudgeon, P. (2017). Some improvements in confidence intervals for standardized regression coefficients. *Psychometrika*, 82(4):928–951; https://doi.org/10.1007/s11336–017–9563–z.

Fraley, R. C. and Vazire, S. (2014). The n-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE*, 9(10):e109019.

Fritz, C. O., Morris, P. E., and Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141(1):2; https://doi.org/10.1037/a0024338.

Gatsonis, C. and Sampson, A. R. (1989). Multiple correlation: exact power and sample size calculations. *Psychological Bulletin*, 106(3):516.

Gelman, A. et al. (2005). Analysis of variance – why it is more important than ever. *The Annals of Statistics*, 33(1):1–53.

Hartung, J., Cottrell, J. E., and Giffin, J. P. (1983). Absence of evidence is not evidence of absence. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 58(3):298–299.

Hättenschwiler, N., Merks, S., Sterchi, Y., and Schwaninger, A. (2019). Traditional visual search versus x-ray image inspection in students and professionals: Are the same visual-cognitive abilities needed? *Frontiers in Psychology*, 10:525; https://doi.org/10.3389/fpsyg.2019.00525.

Hauck, W. W. and Anderson, S. (1986). A proposal for interpreting and reporting negative studies. *Statistics in Medicine*, 5(3):203–209.

Hurvich, C. M. and Tsai, C. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, 44(3):214–217.

Jan, S.-L. and Shieh, G. (2014). Sample size determinations for Welch's test in one-way heteroscedastic anova. *British Journal of Mathematical and Statistical Psychology*, 67(1):72–93; https://doi.org/10.1111/bmsp.12006.

Jan, S.-L. and Shieh, G. (2019). On the extended welch test for assessing equivalence of standardized means. *Statistics in Biopharmaceutical Research*, pages 1–8.

Jeffreys, H. (1961). *The theory of Probability*. OUP Oxford.

Kelley, K. et al. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, 20(8):1–24.

Koh, A. and Cribbie, R. (2013). Robust tests of equivalence for k independent groups. *British Journal of Mathematical and Statistical Psychology*, 66(3):426–434; https://doi.org/10.1111/j.2044–8317.2012.02056.x.

Kruschke, J. K. and Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, 25(1):178–206.

Kühberger, A., Fritz, A., and Scherndl, T. (2014). Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. *PLoS ONE*, 9(9):e105825.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in Psychology*, 4:863; https://doi.org/10.3389/fpsyg.2013.00863.

Lakens, D., Scheel, A. M., and Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269; https://doi.org/10.1177/2515245918770963.

Levy, K. J. (1978). Some empirical power results associated with Welch's robust analysis of variance technique. *Journal of Statistical Computation and Simulation*, 8(1):43–48.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Linde, M. and van Ravenzwaaij, D. (2019). baymedr: An R package for the calculation of Bayes Factors for equivalence, non-inferiority, and superiority designs. *arXiv preprint arXiv:1910.11616*.

Marszalek, J. M., Barber, C., Kohlhart, J., and Cooper, B. H. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, 112(2):331–348.

McCambridge, J., Wilson, A., Attia, J., Weaver, N., and Kypri, K. (2019). Randomized trial seeking to induce the Hawthorne effect found no evidence for any effect on self-reported alcohol consumption online. *Journal of Clinical Epidemiology*, 108:102–109; https://doi.org/10.1016/j.jclinepi.2018.11.016.

Morey, R. D., Rouder, J. N., Jamil, T., and Morey, M. R. D. (2015). Package 'BayesFactor'. *URL http://cran/r-projectorg/web/packages/BayesFactor/BayesFactor*.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.

Nagelkerke, N. J. et al. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.

Nakagawa, S. and Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, 4(2):133–142.

Ohtani, K. (2000). Bootstrapping R2 and adjusted R2 in regression analysis. *Economic Modelling*, 17(4):473–483.

Ohtani, K. and Tanizaki, H. (2004). Exact distributions of R2 and adjusted R2 in

a linear regression model with multivariate t error terms. *Journal of the Japan Statistical Society*, 34(1):101–109.

Okada, K. (2013). Is omega squared less biased? a comparison of three major effect size indices in one-way anova. *Behaviormetrika*, 40(2):129–147; http://dx.doi.org/10.2333/bhmk.40.129.

Pallmann, P. and Jaki, T. (2017). Simultaneous confidence regions for multivariate bioequivalence. *Statistics in Medicine*, 36(29):4585–4603; https://doi.org/10.1002/sim.7446.

Parkhurst, D. F. (2001). Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation equivalence tests improve the logic of significance testing when demonstrating similarity is important, and reverse tests can help show that failure to reject a null hypothesis does not support that hypothesis. *Bioscience*, 51(12):1051–1057.

Plonsky, L. and Oswald, F. L. (2017). Multiple regression as a flexible alternative to anova in l2 research. *Studies in Second Language Acquisition*, 39(3):579–592; https://doi.org/10.1017/S0272263116000231.

Pocock, S. J. and Stone, G. W. (2016). The primary outcome fails -what next? *New England Journal of Medicine*, 375(9):861–870.

Quertemont, E. (2011). How to statistically show the absence of an effect. *Psychologica Belgica*, 51(2):109–127; http://doi.org/10.5334/pb–51–2–109.

Rose, E. M., Mathew, T., Coss, D. A., Lohr, B., and Omland, K. E. (2018). A new statistical method to test equivalence: an application in male and female eastern bluebird song. *Animal Behaviour*, 145:77–85; https://doi.org/10.1016/j.anbehav.2018.09.004.

Rouder, J. N. and Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47(6):877–903; DOI: 10.1080/00273171.2012.734737.

Rusticus, S. A. and Lovato, C. Y. (2011). Applying tests of equivalence for multiple

group comparisons: Demonstration of the confidence interval approach. *Practical Assessment, Research & Evaluation*, 16.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney U test. *Behavioral Ecology*, 17(4):688–690.

Schönbrodt, F. D. and Wagenmakers, E.-J. (2016). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, pages 1–15.

Stand, J. (2000). The "Hawthorne effect" -what did the original Hawthorne studies actually show. *The Scandinavian Journal of Work, Environment & Health*, 26(4):363–367.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., and Van Der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on bem (2011).

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. Chapman and Hall/CRC.

Wellek, S. (2017). A critical evaluation of the current "p-value controversy". *Biometrical Journal*.

Yang, L. and Tsiatis, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321; https://doi.org/10.1198/000313001753272466.

Zimmerman, D. W. (2004a). Conditional probabilities of rejecting h 0 by pooled and separate-variances t tests given heterogeneity of sample variances. *Communications in Statistics-Simulation and Computation*, 33(1):69–81.

Zimmerman, D. W. (2004b). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57(1):173–181.

Zou, K. H., Tuncali, K., and Silverman, S. G. (2003). Correlation and simple linear regression. *Radiology*, 227(3):617–628.