Meta-Psychology, 2024, vol 8, MP.2020.2554
https://doi.org/10.15626/MP.2020.2554
Article type: Original Article

Open data: Yes
Open materials: Not Applicable
Open Code: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Daniel Lakens
Reviewed by: Brydges, C., Field, S.M., Hildegard, J.
Analysis reproduced by: Jens Fust
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/FYUBE

Bayesian Evaluation of Replication Studies

Hidde J. Leplaa[1], Charlotte Rietbergen[1], and Herbert Hoijtink[1]
[1]Department of Methodology and Statistics, Utrecht University

In this paper a method is proposed to determine whether the result from an original study is corroborated in a replication study. The paper is illustrated using two replication studies and the corresponding original studies from the Reproducibility Project: Psychology by the Open Science Collaboration. This method emphasizes the need to determine what one wants to replicate from the original paper. This can be done by translating the research hypotheses formulated in the introduction into informative hypotheses, or, by translating the results into interval hypotheses. The Bayes factor will be used to determine whether the hypotheses resulting from the original study are corroborated by the replication study. Our method to assess the successfulness of replication will better fit the needs and desires of researchers in fields that use replication studies.

Keywords: Bayes factor, Informative Hypothesis, Replication Crisis, Replication Study

Introduction

People will walk more slowly when primed with words related to being old was the result of a study by Bargh et al. (1996). However, the replication study by Doyen et al. (2012) failed to corroborate the results. This study became one of the examples of the existence of the so-called replication crisis in the behavioral sciences; results obtained in a study are often not corroborated by its replication. The question is now: what should we believe? How can we rely on results from scientific research in the behavioral sciences if they contradict each other so often? But first: do they actually contradict each other? We will focus on the question how it can be determined whether a replication study supports the earlier study. We will propose two methods to translate an original paper into a hypothesis: either based on the theory, or on the results. Subsequently, the data resulting from the replication study will be used to evaluate these hypotheses with the Bayes factor.

We will illustrate this paper with two studies from the Open Science Collaboration (OSC) Reproducibility Project: Psychology (OSC, 2012, 2015). For each of 100 psychological experiments, an exact (Hüffmeier et al., 2016) replication study was executed by researchers from the OSC, using a procedure as similar as possible to the original study and indicating in detail any methodological differences between the original study and the replication study. Furthermore, the sample size for the replication study was determined based on a power analysis. The statistical analyses in the replication studies were the same ones as used in the original studies. The goal was to evaluate whether the results of the replication study corroborated the original study; which would mean it was a successful replication. In this project, only 36% to 47% of the original studies was successfully replicated. Since the OSC consists mainly of studies comparing group means, our focus will be on ANOVAs, although the ideas can easily be translated to other statistical models.

It is open for debate how many studies within the OSC (2012, 2015) were replicated. This is due to the fact that different methods were used to determine the success of the replications. Throughout the literature, different kinds of methods are used to evaluate replication studies. Anderson and Maxwell (2016) identify six different goals when evaluating successfulness of replication studies, and present for each goal a suitable method to analyse it: *significance-based*, *null effect inferring*, *effect size estimating*, *meta-analysis*, *assessing inconsistency*, and *assessing consistency*. We will discuss these methods, along with some of the critiques. Both the third, *effect size estimation*, and fourth, *meta-analysis*, do not focus on testing whether a replication study corroborates the results of the original study. Therefore we will not discuss these methods. Then, we will introduce our methods as an additional and innovative option because in contrast to the other approaches we formulate the replication hypotheses explicitly based on the theory or results presented in the original paper.

The significance-based method was done by testing whether the effect observed in both the original and the replication study was significant. It is also referred to as *vote counting* (Anderson & Maxwell, 2016; Simonsohn, 2015; Zondervan-Zwijnenburg et al., 2019). If

the statistical tests provide the same answer (e.g. both significant or both non-significant), and the direction of the effect is the same, the study is considered to be replicated. The problem with this approach is that while an original study with a p-value of .04 is not considered to be replicated if the replication study finds a p-value of .06, the same original study is replicated when the replication study finds a p-value of $< .001$. Interestingly, the difference between a p-value of .04 and .06 is not necessarily significant itself (Gelman & Stern, 2006) For further research in line with the significance-based method we refer to Etz and Vandekerckhove (2016), Ly et al. (2018), Marsman et al. (2017), Simonsohn (2015), and Verhagen and Wagenmakers (2014).

The second method discussed in Anderson and Maxwell (2016) was *null effect inferring*. In this method, usually called equivalence testing, the null hypothesis that there is a relevant effect is tested against the alternative hypothesis, that there is not a relevant effect. This is done by establishing a "region of indifference" (Anderson & Maxwell, 2016). This region identifies which results are essentially zero, assuming that finding an effect of exactly zero is nearly impossible. If then the 90%-confidence interval (90%-CI) around the effect found in the replication falls completely in this region of indifference, the hypothesis that there is a relevant effect has to be rejected in favor of the hypothesis that there is not a relevant effect. If it (partially) lies outside this region, the hypothesis that there is relevant effect can not be rejected. An issue with this method is that it is difficult to determine the width of the region of indifference. The interested reader is referred to Simonsohn (2015) who specifies such an interval using a Cohen's d of .1, however, without providing an argument for this specific interval. Furthermore, the results of the original study are disregarded when deciding to accept or reject the null hypothesis. That is, the region of indifference is determined without taking into account the data and results from the original study, in the next method the original study does have a role when determining the hypotheses of interest.

Assessing inconsistency is an answer to the critique that the difference between significant and non-significant is often not significant itself (Gelman & Stern, 2006). It is a test of the heterogeneity of the effect sizes resulting from the original and replication study. It is performed by calculating a confidence interval around the difference in effect sizes. It is a sophisticated method to assess the similarity between the original and replication study. Since only one effect size per study can be included in the analysis at once the method of *Assessing inconsistency* only properly works for t-tests, which makes it somewhat limited. Our method will extend on this method but will be suitable for ANOVAs as well. Another application of this method can be found in Patil et al. (2016), though their method is limited to a confidence interval for the original study. Furthermore, as Morey and Lakens (n.d.) also highlighted, their method requires well-powered original and replication studies. The latter is not unique to Patil et al.'s approach, and will further be discussed later in this paper.

Last Anderson and Maxwell (2016) identify the method called *assessing consistency*. This method combines the methods two (equivalence testing) and five (assessing inconsistency): they test if the CI around the difference in the effect sizes falls within a subjectively determined region of indifference. As with the second method: the width of the interval is subjectively determined. But, as Anderson and Maxwell (2016) mention themselves: only with very large sample sizes the CI will be small enough to reasonably expect it to fit within the region of indifference. In all applications without very large sample sizes, this method will never confirm that the effect sizes are equal.

Three issues apply to both the first two and the last two methods identified by Anderson and Maxwell (2016). First, all methods result in a dichotomous interpretation of the test results: either the results from the original study are replicated or they are not. However, although this is often ignored, there are error probabilities associated with these decisions: if the null hypothesis is rejected there is usually a Type I error of 5% probability that this is incorrect, and if the null hypothesis is not rejected, there is an unknown (because the population effect size is not known) Type II error probability that the null hypothesis is incorrectly not rejected. Second, the methods test just one hypothesis at a time. As we will show in this paper, sometimes researchers want to test multiple competing hypotheses. This can be the case in, for example, a two-way ANOVA, where the goal may be to determine if none, one, or both possible main effects are present (as is the case in the study by Janssen et al. (2008). Third and last, all methods test a (interval) hypothesis against the corresponding alternative hypothesis. As will be shown in this paper our approach also allows for a class of more flexible hypotheses that are called informative hypotheses (Hoijtink, 2012).

The approaches proposed in this paper is a new and innovative addition to the approaches identified by Anderson and Maxwell (2016). First, we use the Bayes factor (BF) (Hoijtink, Gu, et al., 2019b; Kass & Raftery, 1995) which replaces dichotomous decisions by a quantification of the support in the data for each of the hypotheses involved. Use of support as quantified by the Bayes factor is elaborated by Morey et al. (2016). It is an alternative for Popperian falsification (Popper, 1963)

called logical positivism. The interested reader is referred to Hawthorne (2021) for an elaboration of logical positivism, that is, the philosophical foundation of using the BF as a measure of support for the hypotheses of interest. Second, we allow for the evaluation of multiple hypotheses at once, with the additional benefit that the methods can be applied to ANOVAs and *t*-tests while the other approaches are mostly tailored to the *t*-test. Third, we force the replication researcher to make an informed decision about the formulation of the replication hypotheses.

The BF was already proposed by Anderson and Maxwell (2016) to evaluate the results of replication studies. As will be elaborated later, the Bayes factor measures the relative support in the data for two or more competing hypotheses. The BF has been used previously to evaluate replication studies. Some of these studies used the ANOVA or *t*-test null hypotheses (Etz & Vandekerckhove, 2016; Harms, 2018; Ly et al., 2018; Marsman et al., 2017) versus a one- or two-tailed alternative hypothesis. Other studies using the BF only focused on *t*-tests (Field et al., 2019; Marsman et al., 2017; Van Aert & Van Assen, 2017).

Cho and Abe (2013) addressed the fact that researchers do not always test the correct hypothesis. Often, this involves testing the null hypothesis, which was previously criticized by Gigerenzer (2004). We argue that oftentimes retesting the null hypothesis is not in line with the goals of the replication researchers, who want to determine if the replication study corroborates the results from the earlier study. To test more relevant hypotheses, we allow the replication researchers to evaluate replication hypotheses based on either the introduction of the original paper (subsequently called the "theory based method") or the results of the original paper (subsequently called the "results based method"). This is not unique to our approach, other examples can be found in Simonsohn (2015), who use the results of the original study to determine the effect size to be evaluated in a replication study. Another example, although not specific to the evaluation of replication studies, is replacing the null hypothesis with an interval hypothesis, as is done in equivalence testing (Lakens et al., 2018).

In our methods we will evaluate other kinds of hypotheses than used in these articles. The theory based method translates the introduction section from an original paper into an informative hypothesis or hypotheses (Hoijtink, 2012, e.g. $H_t : \mu_3 < \mu_2 < \mu_1$), whereas the results based method uses the results section to formulate interval hypotheses (Lakens et al., 2018; Zondervan-Zwijnenburg et al., 2019, e.g. $H_r : 5.28 < \mu_1 < 5.94$ & $4.90 < \mu_2 < 5.56$). After data is collected the BF is calculated to determine the relative support in the data for the replication hypothesis (the informative or interval hypothesis) versus its complement. This renders an innovative method that uses informative hypotheses and the BF to evaluate replication studies. Our approach explicitly addresses the fact that researchers do not always test the correct hypothesis (Cho & Abe, 2013) and therefore is a remedy against the "null-ritual", which was previously criticized by Gigerenzer (2004). Our paper is also in line with Dienes (2014) who discusses Bayesian interpretations of non-significant results (however, not in the context of replication studies). To address a similar issue, we step away from using the null-hypothesis, instead we use informative and interval hypotheses, and put this to work in the context of replication studies.

This paper is structured as follows: our method is illustrated using one of the replication studies from the OSC (2012, 2015). This study will be introduced in the next section. Following this introduction, we will explain and demonstrate our methods, where an original paper is translated to hypotheses, from start to finish. Subsequently, a second example employing more than one hypothesis of interest will be presented. This paper is concluded with a discussion in which we will reflect on the possibility to synthesize information from different studies, similar to methods three and four of Anderson and Maxwell (2016).

The running example

The running example used in this paper is a study on the perceptions of closeness toward one's family members and hometown (Williams & Bargh, 2008). In Study 4 of this paper, participants were instructed to mark off two points on a Cartesian coordinate plan: these could be close to the origin (Group 1 - Closeness), some distance from the origin (Group 2 - Intermediate), or far from the origin (Group 3 - Distance). Participants then rated the strength of their bonds to their siblings, their parents, and their hometown. All questions were answered on a scale ranging from 1 (*not at all strong*) to 7 (*extremely strong*) and averaged to create a total score. The hypothesis was that marking points more distant from the origin would result in perceiving the closeness toward one's family members and hometown as less strong. The results from this study are summarized in Table 1, which contains more information and will be referred to throughout this paper. Williams and Bargh (2008) concluded that their expectations were matched by their results. This can be seen from the *p*-value of .01, which indicates that the three means are not equal, and an ordering of the means as hypothesized.

Joy-Gaba et al. (2012) executed a direct replication of the study by Williams and Bargh (2008). Their proce-

dures were as similar as possible, using the same materials in the replication study as were used in the original study. The results from the replication study, also presented in Table 1, show no significant differences between the groups. Using the procedure of *vote counting* it can be concluded that the results were not successfully replicated: the *p*-value is not below the threshold of .05.

What should be replicated

As elaborated in the introduction, our method to evaluate replication studies determines the support in the replication data for one ore more hypotheses resulting from the original study. Two methods to obtain these hypotheses will be presented in this section. In the first, the hypotheses are extracted from the theoretical framework as presented in an introduction of the original paper. In the second, hypotheses are formulated based on the information in the results section of the original paper. An overview of the steps is given in Table 2.

Method 1 - Theory based method

To construct hypotheses from the introduction of the original study a six step procedure will be used that is based on qualitative research methods (e.g. Charmaz, 2006; Glaser, 1978).

Step 1.1 - Reading the introduction

Read the introduction section of the original paper. The introduction typically contains the theoretical background for the study, and results in the formulation of theory-based hypotheses that will be tested in the study. There is no output from this step.

Step 1.2 - Coding of statements in the introduction

Those statements are coded that specify theories or expectations with respect to the outcomes of the study. Coding is a qualitative (Boeije, 2010; Charmaz, 2006; Glaser, 1978) manner of data analysis. Typically, coding consists of two activities: segmenting and labeling (textual) data. For our purposes, the labelling part is not that relevant, because the only goal is to localize statements that can be labeled "theory or expectation with respect to the outcomes of the study", that is, for each statement selected the label is the same. The statements of interest are often indicated with words such as "hypothes*", "expect*", and "argue*" [1].

This step gives as output one or more statements. In the running example two statements were coded:

1. "Accordingly, we argue that a primitive understanding of physical distance is the foundation for the later-developed concept of psychological distance, given humans' pervasive tendency to conceptualize the mental world by analogy to the physical world" (Williams & Bargh, 2008, p.2)

2. "We hypothesized that participants primed with distance would report weaker attachments to their family members and hometown, compared with participants primed with closeness" (Williams & Bargh, 2008, p.6)

Step 1.3 - Reading the methods and the results section

Read the results section of the original paper. This section elaborates which analyses are to be executed. No output results from this step.

Step 1.4 - Coding of statements in the methods and the results section

Statements are coded that indicate what is actually tested. In theory, this can be statements in the methods section explaining which tests will be conducted to test the hypotheses of interest. However, after checking 25 papers from the OSC Reproducibility Project: Psychology to find examples of ways to present this information in the methods section, we had to conclude that this information is never, or barely ever, presented in the methods section. Therefore only the results section is used for the following steps, since here the relevant information is always present. It remains necessary to read the methods section, to understand how the variables used in a study are operationalized.

In the results section the outcomes of the tested hypotheses are presented. However, note that, these will only be coded according to which hypothesis has been evaluated with which test, not to clarify the results of that test. The statements of interest may be indicated with words such as: "we have tested (...)", "we used/conducted (...) analysis/test (...)", and "to test the hypothesis (...)". Also, sentences indicating particular tests are of interest, e.g. *t*-test, ANOVA, and regression. For the running example the following statements were coded in the results section:

1. "Next, we conducted a planned contrast analysis using weights of -1, 0, and 1 for a linear contrast, and -1, +2, -1 for a quadratic contrast, for the Closeness, Intermediate, and Distance conditions, respectively. These contrast weights allowed us to test the specific hypothesis that participants in

[1] *= different extensions are possible, e.g. hypothes**is**, hypothes**ize**, hypothes**es**

Table 1

Descriptive statistics and results of the original study of Williams and Bargh (2008) and its replication by Joy-Gaba et al. (2012).

	Original study by Williams & Bargh (2008)			Replication study by Joy-Gaba et al. (2012)		
	Group 1 Closeness	Group 2 Intermediate	Group 3 Distance	Group 1 Closeness	Group 2 Intermediate	Group 3 Distance
M (SD)	5.61 (0.90)	5.23 (0.90)	4.86 (0.90)	5.44 (0.83)	5.31 (1.07)	5.31 (1.15)
95% CI	(5.28, 5.94)	(4.90, 5.56)	(4.53, 5.19)	(5.19, 5.69)	(4.99, 5.63)	(4.94, 5.68)
n	28	28	28	44	43	38
M_{1-2} (95% CI)	0.38 (-0.09, 0.85)			0.13 (-0.27, 0.53)		
M_{2-3} (95% CI)	0.37 (-0.10, 0.84)			0.00 (-0.48, 0.48)		
$H_0 : \mu_1 = \mu_2 = \mu_3$	$p = .01$, $\eta^2 = .11$			$p = .79$, $\eta^2 = .004$		
$H_{t_s}{}^* : \mu_3 < \mu_2 < \mu_1$				$BF_{t_s c} = 2.10$		
$H_{t_{min1}} : \mu_3 + 0.21 < \mu_2 + 0.10 < \mu_1$				$BF_{t_{min1} c} = 0.57$		
$H_{t_{min2}} : \mu_3 + 0.41 < \mu_2 + 0.19 < \mu_1$				$BF_{t_{min2} c} = 0.11$		
$H_{r_M} : 5.28 < \mu_1 < 5.94$ &				$BF_{r_M c} = 1.92$		
$4.90 < \mu_2 < 5.56$ & $4.53 < \mu_3 < 5.19$						
$H_{r_D} : -0.09 < \mu_1 - \mu_2 < 0.85$ &				$BF_{r_D c} = 4.44$		
$-0.10 < \mu_2 - \mu_3 < 0.84$						

Note that, Williams and Bargh (2008) do not present their three means, standard deviations and sample sizes. We were unable to retrieve the original data, therefore numbers presented in the table were reconstructed using the information presented in the paper assuming equal within groups variances and assuming that the mean of the second group was located on an equal distance of the other two means. This is very likely not in line with the original data, but allows us to illustrate our approach using a real original study.

*= In the introduction section Williams and Bargh only explicitly discuss the groups Closeness and Distance, but in line with the theory underlying distance ques, we assumed that the effect of "Intermediate" is between that of "Distance" and "Closeness".

the Distance condition enjoyed the excerpt more than those in the Intermediate condition, who in turn enjoyed it more than those in the Closeness condition." (Williams & Bargh, 2008, p.4)

2. "An ANOVA revealed that the three spatial-prime groups differed significantly in the reported strength of their bonds to their family and hometown, (...)" (Williams & Bargh, 2008, p.7)

This step results in one or more coded statements.

Step 1.5 - Selecting the statements from the introduction that are actually tested

The statements coded in Step 1.4 can be used to verify which statements coded in Step 1.2 are actually tested. For the running example, as can be seen from the second statement coded in Step 1.4, only the second statement coded in Step 1.2 ("We hypothesized that participants primed with distance would report weaker attachments to their family members and hometown, compared with participants primed with closeness") was tested. The result of Step 1.5 is a number of statements.

Table 2

Overview and short explanation of the steps.

Theory based method	**Results based method**
1.1 - Reading the introduction	**2.1 - Reading the results section**
Simply read through the section to know what is being studied	Simply read through the section to know what is being studied
1.2 - Coding statements in the introduction	**2.2 - Extracting the results**
Indicate the prior expectations of the researchers	Extract the results and indicate
1.3 - Reading the methods and the results section	what the results of the conducted tests are
Simply read through the section to know what tests were conducted	**2.3 - Translating results to interval hypotheses**
1.4 - Coding of statements in the	Calculate the desired interval hypotheses based on the results
methods and the results section	
Select the statements that report the results of the tests	
1.5 - Selecting the statements from the	
introduction that are actually tested	
Select the statements from Step 1.2 that are tested	
with the statements selected in Step 1.4	
1.6 - Translating statement to informative hypotheses	
Formulate the tested statements from the	
introduction section as testable hypotheses	

Step 1.6 - Translating statements to informative hypotheses

In Step 1.6 informative hypotheses are formulated. In the theory based method only ordered hypotheses are formulated, of which two forms are distinguished: simple ordering and minimal difference hypotheses on (combinations of) parameters.

Informative hypotheses are a formal representation of the expectations a researcher has with respect to the relations between the means in an ANOVA (Hoijtink, 2012, p. 50-51; Klugkist, 2005). Using inequality constraints, order relations between (combinations of) means (μ) can be specified. Three examples are:

$$H_1 : \mu_1 < \mu_2 < \mu_3,$$

$$H_2 : \mu_1 + 0.2 < \mu_2 \ \& \ \mu_1 + 0.5 < \mu_3,$$

and

$$H_3 : \mu_{1A} - \mu_{1B} > \mu_{2A} - \mu_{2B}.$$

H_1 shows a specific ordering of three means, where the mean of Group 1 is smaller than that of Group 2, which in turn is smaller than that of Group 3. H_2 is also an ordering of three means, but now the minimal difference is specified. The mean of Group 1 is at least 0.2 points lower than the mean of Group 2, and 0.5 points lower than the mean of Group 3. H_3 shows a two way analysis of variance, with two factors with levels 1/2 and A/B. Here, it is stated that the difference between the means of Levels A and B is greater when located at Level 1, than when located at Level 2. This is the start of a hypothesis that specifies an interaction effect. To further specify the interaction effect, elements like $\mu_{1A} > \mu_{1B}$ and $\mu_{1A} > \mu_{2A}$ have to be added.

Next to inequality constraints, one could also use equality constraints. Two examples are:

$$H_4 : \mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B},$$

and

$$H_5 : \mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B} \ \& \ \mu_{1A} + \mu_{2A} > \mu_{1B} + \mu_{2B}.$$

H_4 indicates the absence of a main effect of the factor with levels 1/2. No specification has been given in H_4 regarding a main effect of the factor with levels A/B or an interaction effect of those two factors. Next, H_5 is an extension of H_4. Next to the specified absence of a main effect of the factor with levels 1/2, a main effect of the factor with levels A/B is specified here. H_5 hypothesizes that there is main effect, with the subjects located at level A score on average higher than subjects located at level B.

To formulate the informative hypothesis based on the theory section of the paper, first the dependent variable needs to be be determined. Second, it needs to be determined which groups are distinguished and the theory needs to be translated into inequality and or equality restrictions on the group means. Applied to the running example this renders: the variables in the statement ("We hypothesized that participants primed with distance would report weaker attachments to their family members and hometown, compared with participants primed with closeness"; Williams & Bargh, 2008, p.6) are 1) attachment to family members and hometown, and 2) the prime. The first variable, attachment, is measured on a 7-point interval scale. For the prime

variable three groups are distinguished in this study; Distance, Intermediate and Closeness. The groups are clearly ordered in this statement: a group primed with more closeness scores is expected to score on average higher on the dependent variable, strength of the attachment, than the group primed with more distance. Because there is no mention of the size of the difference, an informative hypothesis with an inequality constraint will be formulated:

$$H_{t_s} : \mu_{Distance} < \mu_{Intermediate} < \mu_{Closeness},$$

where, the subscript t denotes that the hypothesis is based on the literature review in the introduction of the paper, and the sub-subscript s indicates a simple ordering hypothesis is formulated. We would like to add another variant of these types of hypotheses. Typically, when discussing differences between two means, people refer to a relevant difference. The hypothesis H_{t_s} leaves room for irrelevant differences, e.g. the mean of the Distance group is 0.01 lower than the mean of the Closeness group. In the second hypothesis based on the introduction section we assume at least a small effect, that is a minimal difference of 0.2 standard deviation (*SD*; Cohen, 1988). There are two ways to interpret this: either 0.2 *SD* difference between adjacent groups or between the groups with the most extreme means. To show the effect of either decision, two hypothesis are formulated:

$$H_{t_{min1}} : \mu_{Distance} + 0.21 < \mu_{Intermediate} + 0.10 < \mu_{Closeness}$$

and

$$H_{t_{min2}} : \mu_{Distance} + 0.41 < \mu_{Intermediate} + 0.19 < \mu_{Closeness}$$

where the sub-subscript *min* indicates that a minimal difference hypothesis is formulated with at least Cohen's *D* equals 0.2 between adjacent groups. The score of 0.10 equals 0.1 times the pooled within group standard deviation of groups Intermediate and Closeness in the replication data. The score of 0.19 equals 0.2 times the pooled within group standard deviation of groups Intermediate and Closeness in the replication data. The scores of 0.21 and 0.41 equal 0.1 and 0.2 times the pooled within group standard deviation of groups Distance and Intermediate on top of the difference between the groups Intermediate and Closeness: i.e. 0.21 is based on 0.0956 (0.1 times pooled within group *SD* for Intermediate and Closeness) + 0.1108 (0.1 times pooled within group *SD* for Distance and Intermediate). The result of this step is an informative hypothesis that can later be evaluated using the data resulting from the replication study. Before we continue with the analysis, we will discuss the results based method.

Method 2 - Results based method

To construct hypotheses based on the outcomes of the original study a three step procedure will be used.

Step 2.1 - Reading the results section

Read the results section of the original paper, which reports the outcomes of the conducted analyses. It might also be useful to read the methods section to get a general idea of how the study was executed. There is no output from this step.

Step 2.2 - Extracting the results

Extract the results from the original paper. The relevant results for this method are the mean and its standard error for each group. If the standard error is not reported it can be computed using the standard deviation and sample size for each of the groups. Those statistics are often included in a table with descriptive statistics in the results section. In the running example not all the needed statistics were reported. The available statements that were used to determine the results reported in Table 1 are:

1. "An ANOVA revealed that the three spatial-prime groups differed significantly in the reported strength of their bonds to their family and hometown, $F(2, 81) = 4.97$, $p_{rep} = .95$, $\eta^2 = .11$." Note that, this corresponds to the *p*-value of .01 as reported in Table 1.

2. "(...) the linear contrast showed that participants primed with distance reported weaker bonds to their family and hometown ($M = 4.86$), compared with participants primed with closeness ($M = 5.61$), $t(81) = -2.86$, $p_{rep} = .96$." Note that, this corresponds to the *p*-value of .01 as reported in Table 1.

Step 2.3 - Translating results to interval hypotheses

With interval hypotheses it can be tested whether one or more means are within our outside of a certain interval of values (Zondervan-Zwijnenburg et al., 2019). Examples of interval hypotheses are:

$$H_6 = 0.2 < \mu_1 < 0.4,$$

$$H_7 = -0.1 < \mu_1 < 0.7 \ \& \ 1.1 < \mu_2 < 2.4,$$

and

$$H_8 = -0.6 < \mu_1 - \mu_2 < 0.1 \ \& \ 0.7 < \mu_2 - \mu_3 < 1.2.$$

In H_6 it is stated that the mean of Group 1 is within the range 0.2 to 0.4. Whereas H_7 states that the mean of Group 1 is somewhere in the range -0.1 through 0.7, and the mean of Group 2 is in the interval 1.1 to 2.4. Lastly, H_8 states that the difference between the means of Groups 1 and 2 is somewhere in the range of -0.6 (μ_2 is considerably larger) to 0.1 (μ_1 is slightly larger), and the difference between the means of Groups 2 and 3 is in the range of 0.7 to 1.2 (μ_2 is larger). The end points of the interval should be chosen such that they represent the uncertainty in the estimates of the means obtained from the original study. As can be seen in Equations 1 and 2 the lower bound (LB) of the interval is the lower bound of the traditional 95% confidence interval for the mean or the difference between two means, and the upper bound (UB), Equations 3 and 4, is chosen analogously:

$$LB_{mean} = M - 1.96 * \frac{SD}{\sqrt{n}} \tag{1}$$

$$LB_{difference} = M_1 - M_2 - 1.96 * \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}} \tag{2}$$

$$UB_{mean} = M + 1.96 * \frac{SD}{\sqrt{n}} \tag{3}$$

$$UB_{difference} = M_1 - M_2 + 1.96 * \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}, \tag{4}$$

where all statistics are derived from the original study.

Due to publication bias (Rosenthal, 1979) $M_1 - M_2$ is possibly an overestimation of the true effect. If this is indeed the case, $M_1 - M_2$ obtained from the replication data will probably be (much) smaller and not contained in the intervals specified in the hypotheses. Therefore, the support in the replication data for the replication hypotheses will be small, which is, in this case, the desired result. If the interval in the replication hypothesis becomes very narrow, this would be the case when the sample size in the original study was large, two things can happen. Either the result from the replication study falls within the interval, which would lead to strong support for the replication hypothesis as is desired. Or, if bias was present in the original study, the results from the replication study falls outside of the interval which would lead to a strong rejection of the replication hypothesis, which is also desired.

To formulate an interval hypothesis, a similar procedure should be followed as for the theory based method. Start with determining the dependent variable. Then, determine which groups are distinguished. Finally, the confidence interval for each group can be determined. In the running example, three means were reported. For each of those means an interval has to be specified. Using the descriptive statistics presented in Table 1 and Equations 1 and 3 results in:

$$H_{r_M} : 5.28 < \mu_{Closeness} < 5.94 \ \& \ 4.90 < \mu_{Intermediate} < 5.56 \ \& \ 4.53 < \mu_{Distance} < 5.19,$$

where the subscript r indicates the hypothesis is formulated based on the results, and the sub-subscript M indicates it is an interval hypothesis based on the means. There is another possibility however. Instead of focusing on the means, one could also focus on the difference between means. Often, researchers are not necessarily interested in the means themselves, but rather on the respective position of one mean related to another. Therefore, interval hypotheses could also be formulated such that they specify in what region the difference between two means fall. For this study it results in the following hypothesis:

$$H_{r_D} : -0.09 < \mu_{Closeness} - \mu_{Intermediate} < 0.85 \ \& \ -0.10 < \mu_{Intermediate} - \mu_{Distance} < 0.84,$$

where the sub-subscript D indicates the interval hypothesis is based on the difference between the means. In hypothesis H_{r_D} it is specified that the difference between the means of the groups Closeness and Intermediate should be between -0.09 (mean of the Intermediate group is somewhat higher) and 0.85 (mean of Closeness is considerably higher), and the difference between the means of the groups Intermediate and Distance is in the range of -0.10 to 0.84. This interval is based on the 95%-CI calculated around the difference of these means in the original study. Furthermore, the difference between the means of the Intermediate and Distance group is also specified in hypothesis H_{r_D}.

At this point, the replication hypotheses are formulated in the form of informative hypotheses in the theory based method or in the form of interval hypotheses in the results based method. In the following section we discuss how these hypotheses can be evaluated.

An important property of the original study

An important note needs to be made. When using the original paper to formulate hypotheses, the sample size of that original study is of high influence on usefulness of the replication (Anderson & Maxwell, 2016; Anderson et al., 2017). Note that this only influences the potential usefulness of replications conducting according to the results based method. These problems are not

present in the theory based method, as we will discuss below. We specify two situations that both seem problematic: a large and a small sample size in the original study. This phenomenon is touched upon in the previous subsection (and is also explored in Simonsohn, 2015), but will be discussed more elaborate here.

First, a large sample size might seem problematic. A large sample size leads to small intervals, that is, rather precise estimates. Two situations can arise. Either the estimate is correct or there was bias in the original study. While this may seem problematic, it actually is not. In case the estimate is unbiased, the replicated (difference) in means will be close to the original (difference) in means leading to strong support for the replication hypothesis. However, if bias was present in the original study, the (difference in) means in the replication study will be rather different from the difference specified in the replication hypothesis. This will lead to a rather strong rejection of the replication hypothesis. Both situations give the desired result: either strong support for the replication hypothesis in the absence of bias, or strong support against the replication hypothesis in the presence of bias. Concluding: a large sample size, and bias, are no problem.

Second, a small sample seems problematic. A small sample size leads to relatively big intervals. The bigger an interval, the higher the chances are that the replication hypothesis receives support. One should wonder if it is useful to conduct the replication study when the replication hypothesis is vague and non informative. The result based method should not be used for original studies with too small sample sizes. A critical assessment of the replication hypothesis is needed: do you think it is realistic to reject the replication hypothesis based on the results? If this is not the case, different kinds of hypotheses could be formulated. For instance, the informative hypothesis from the theory based method could be used, even based on the results. That is, the ordering could be specified based on the ordering in the results of the original study. Note that for the theory based method the small sample issue does not arise, since the hypotheses are formulated without involving the sample sizes or reported means and standard deviations from the original study.

To summarize, the theory based method could be used regardless of the properties of the original study. The result based method can be used when the original study has a large sample size, regardless of possible bias. In the case of a small sample in the original study, it is better to avoid the result based method and to use the theory based method. The flexibility in formulating hypotheses ensures that a proper evaluation of the corroboration of the results should be possible.

Conducting the replication study

After determining the hypotheses that will be tested, the replication study can be executed. This includes the set-up of the study, obtaining the participants, and collecting the data. The methodological set-up of the replication study is not within the scope of this paper. For the decisions to be made for the set-up, we refer to articles such as Brandt et al. (2014), Hüffmeier et al. (2016), and Wilson (2016). The three relevant issues to be discussed here are the competing hypothesis, the interpretation of the Bayes factor (BF), and Bayesian updating.

The competing hypothesis

The goal of the replication study is to use new data to determine the relative support for the replication hypothesis versus its complement: "not the replication hypothesis". This enables answering the question "Are the results of the original study replicated or not". The complement of an informative hypothesis specifying two group means is easy to understand: the complement of $\mu_1 < \mu_2$ is simply put $\mu_1 \geq \mu_2$. For hypotheses with more groups and interval hypotheses it is harder to capture the complement. Therefore, the notation "not H_t" is used. For instance, in the theory based method the replication hypothesis specified that the mean of the Distance-group was smaller than the mean of the Intermediate-group which in its turn is smaller than the mean of the Closeness-group (H_{t_s}: $\mu_{Distance} < \mu_{Intermediate} < \mu_{Closeness}$). The competing hypothesis contains every situation where this is not the case (e.g. $\mu_{Intermediate} < \mu_{Distance} < \mu_{Closeness}$, but also $\mu_{Closeness} < \mu_{Intermediate} < \mu_{Distance}$ and more), H_c: "not H_{t_s}". Also for the other hypotheses in the running example H_c equals "not the replication hypothesis".

Comparing the replication hypothesis with its complement

The BF will be used to determine the relative support in the data for the replication hypothesis versus its complement. See for a general introduction of the BF Kass and Raftery (1995), and for a tutorial focused on hypotheses evaluation Hoijtink, Mulder, et al. (2019). Here, the Bayes factor as implemented in the `R` package `bain` will be used (for the statistical background the interested reader is referred to Gu et al., 2018; Hoijtink, Gu, et al., 2019b; and Hoijtink, Mulder, et al., 2019). We now provide a rather accessible account of this BF. For a more thorough explanation and elaboration of how `bain` calculates the BF, the interested reader is referred to . This BF is calculated by evaluating the fit and the complexity of each hypothesis:

$$BF_{t_s c} = \frac{f_t}{c_t} \times \frac{1 - c_t}{1 - f_t}, \quad (5)$$

where f_t denotes the fit of H_t, c_t denotes the complexity of H_t and $1-f_t$ and $1-c_t$ denote the fit and complexity of the complement, respectively. When informative and interval hypotheses are evaluated both the fit and the complexity are probabilities. A really badly fitting hypothesis has a fit close to zero, and a very good fitting hypothesis has a fit close to one. A very specific hypothesis has a complexity close to zero, and a hypotheses that is not very specific has a complexity close to one. Note that, BF_{rc} is obtained if the subscripts t in Equation 5 are replaced by r.

The BF is a measure of the relative support in the data for H_r and H_c. If, for example, $BF_{rc} = 9$, there is 9 times more support in the data for H_r than for H_c. If $BF_{rc} = 0.2$ there is 5 times more support for H_c than for H_r. As can be seen in Equation 5 support is quantified using fit (which should be good) and complexity (which should be small because more specific hypotheses allow for better predictions). The statistical elaboration of fit and complexity can be found in Gu et al. (2018) and Hoijtink, Mulder, et al. (2019). Here only two short and intuitive illustrations will be given. The fit of $H_{t_s} : \mu_1 > \mu_2 > \mu_3$ is good if $M_1 = 6, M_2 = 3, M_3 = 1$ because the sample means are in agreement with the order restrictions specified by the hypothesis, and bad if $M_1 = 5, M_2 = 6, M_3 = 9$. The complexity of $H_{t_s} : \mu_1 > \mu_2 > \mu_3$ equals 1/6 because one of six possible ordering of three means is specified. The complexity of $H_{t'_s} : \mu_1 > (\mu_2, \mu_3)$ equals 2/6 because two orderings are specified ($\mu_1 > \mu_2 > \mu_3$ and $\mu_1 > \mu_3 > \mu_2$). Thus, H_{t_s} is more specific than $H_{t'_s}$ and therefore it has a smaller complexity.

To compute the complexity a parameter called "fraction" has to be set. This parameter determines the weight of the complexity in the computation of the Bayes factor. The Bayes factor for the evaluation of an informative hypothesis specified using only inequality constraints versus its complement is independent of the choice of this fraction (Hoijtink, Mulder, et al., 2019; Mulder, 2014). However, this is not the case for interval hypotheses or informative hypotheses specified using at least one equality constraint. The default setting in `bain` (`fraction = 1`) is chosen such that the Bayes factor tends to prefer the interval hypothesis or an informative hypothesis using at least one equality constraint over its complement unless there is substantial evidence in the data against the interval or informative hypothesis. There is a good reason to prefer the interval hypothesis or informative hypothesis specified with at least one equality constraint (since these state that there is no relevant effect) over its complement (which states that there is a relevant effect). "In an era of heightened awareness of publication bias, sloppy science, and irreplicability of research results, researchers should be conservative, that is, convincing evidence is needed before the alternative hypothesis [there is a (relevant) effect] is preferred over H_0 [or there is no relevant effect]" (Hoijtink, Mulder, et al., 2019, p. 30). However, when the interval hypothesis or the informative hypothesis represents the results of an original study, the reasoning should be the other way around, that is, convincing evidence is needed before the interval or informative hypothesis is accepted over its complement, that is, the Bayes factor should prefer the complement unless there is substantial evidence in the data in favor of the interval or informative hypothesis. This can be achieved in `bain` using `fraction = 3`. The interested reader is also referred to the section Sensitivity Analysis in Hoijtink, Mulder, et al. (2019) where this topic is further elaborated.

Note that, the BF does not lead to a dichotomous decision in favor or against the replication hypothesis, it is a measure of support. If $BF_{rc} = 82$, it is clear that the replication hypothesis is preferred by the data, if $BF_{rc} = 9$, there still is substantial support for the replication hypothesis; if $BF_{rc} = 3.5$, there is some support for the replication hypothesis but its complement can not be disqualified; and, if $BF_{rc} = 1.4$, the replication data provide about equal support for both hypotheses. For a proper interpretation of the BF, general purpose cut-off values are unnecessary and, in fact, should be avoided (see, for example, the discussion of cut-off values in Hoijtink, Mulder, et al., 2019).

Assuming that a priori each of the hypotheses is equally likely, as is usually done, the BF can be translated into so-called posterior model probabilities (PMP) which quantify the certainty of our preference for one of the hypotheses being tested. For example, if $BF_{rc} = 4$, $PMP_r = 4/(1+4) = .80$ and $PMP_c = 1/(1+4) = .20$, which implies that a preference for H_r comes with a Bayesian error probability of .20. If, for example, $BF_{rc} = 11$, $PMP_r = 11/(1+11) = .92$ and $PMP_c = .08$. This implies that in this situation a preference for H_r comes with a Bayesian error probability of .08. The interested reader is referred to Hoijtink, Mulder, et al. (2019) for further elaboration. Note that, in the sequel we will use both BFs and PMPs to quantify the support in the data for the hypotheses entertained.

Determining the desired support

How much support we aim for if the replication hypothesis or its complement is true needs to be determined in advance. Sufficient support need not be es-

pecially strong (e.g.: $BF_{rc} = 4$, or $BF_{rc} = 6$) if the researcher is testing an idea or hypothesis new to their research field and simply wants to know whether it is worthwhile to pursue this theory. However, when established theories are retested and the researcher is thus very certain about the hypothesis of interest, more support would be required (e.g.: $BF_{rc} = 9$, or $BF_{rc} = 11$). Each researcher has to decide for the situation at hand how much support is required. In line with the last lines of the previous section, to avoid the introduction of new cut-off values, general purpose guidelines will not be provided. Please note that the desired support is also needed in the next subsection where we will introduce Bayesian updating. Note furthermore, that the use of support should not lead to dichotomous decisions. What it should lead to is a preference for one of the hypotheses involved, where the strength of this preference is quantified using the BF and PMPs. E.g. if $BF_{rc} = 6$, there is a preference for H_r, $PMP_r = .86$, but H_c can not yet be ignored, since a preference for H_r comes with a Bayesian error probability of $(1 - .86) = .14$. If $BF_{rc} = 9$, there is a stronger preference for H_r, but still H_c can not completely be ignored: a preference for H_r comes with a Bayesian error probability of .10.

For the running example, different decisions could be made per hypothesis. For the theory based method ($H_{t_s} : \mu_{Distance} < \mu_{Intermediate} < \mu_{Closeness}$ and both $H_{t_{min1}}$ and $H_{t_{min2}}$), a support of $BF_{tc} = 3$ or $BF_{ct} = 3$ could suffice. From the introduction of Williams & Bargh (2008) it can be seen that they pursue the evaluation of a new theory, therefore one could see this as a first exploration where some support would already be satisfactory. The hypotheses from the results based method ($H_{r_M} : 5.28 < \mu_{Closeness} < 5.94$ & $4.90 < \mu_{Intermediate} < 5.56$ & $4.53 < \mu_{Distance} < 5.19$ and $H_{r_D} : -0.09 < \mu_{Closeness} - \mu_{Intermediate} < 0.85$ & $-0.10 < \mu_{Intermediate} - \mu_{Distance} < 0.84$) should receive a substantial amount of support. The data in the original study was very convincing with respect to this hypothesis, therefore it would be reasonable to aim for a BF of 10 in favor of either this replication hypothesis or its complement. Note that, the BF values chosen in this paragraph are, in the role of replication researchers, our choices. Other replication researchers may very well come up with different values.

An important property of the replication study

In classical statistics power analyses are often conducted (Cohen, 1988; Erdfelder et al., 1996). Power analysis provides the needed sample size to reach meaningful conclusions for hypotheses formulated in standard forms. For Bayesian statistics there is a very good alternative: Bayesian updating (Hoijtink, Mulder, et al., 2019; Rouder, 2014). Bayesian updating is conducted by sampling additional data until the minimum amount of support in favor of the replication hypothesis or its complement is obtained or the available resources (time and money) to collect additional data have been exhausted. Of course, replication studies should be adequately powered, see for example Morey and Lakens (n.d.) and Simonsohn (2015). Therefore, it is recommended to execute some form of Bayesian design analysis in order to determine whether the resources available are sufficient to obtain a "well-powered" replication study. Bayesian design analysis is under development, the interested reader is referred to Fu (2022) and Schönbrodt and Wagenmakers (2018) for the current state of the art.

Practically updating can be executed by starting with a reasonable sample size, e.g. $n = 20$ per group. After the first round of data collection the BF is determined. If this BF meets the desired support, see the previous subsection, data collection is finished and the research process moves on. If the BF does not meet the desired support for either hypothesis, more data is collected. After each round of data collection (with for instance $n = 20$ per group) the BF is determined. Once it reaches the desired amount of support for either hypothesis the data collection stops. Please note that when the desired BF is relatively small, e.g. $BF_{rc} = 4$ (which corresponds to PMPs of .80 and .20 for H_r and H_c, respectively), the probability that the best hypothesis is incorrectly preferred is still relatively large. On the other hand, if the desired BF is relatively large, e.g. $BF_{rc} = 11$, the probability of an incorrect preference is relatively small, that is, if H_r is preferred the error probability, that is, the PMP associated with H_c, equals .08. Updating cannot be illustrated here, since this paper is based on existing original and replication studies and therefore the sample sizes are fixed. However, we will provide one example where updating is applied using additional hypothetical data and mention for the other hypotheses what decision would be made with regards to updating.

Evaluating the replication hypothesis using the replication data

To calculate BF_{tc} or BF_{rc} the `R` package `bain` (Hoijtink, Mulder, et al., 2019; https://informative-hypotheses.sites.uu.nl/software/bain/) is used. The codes and data used for the analyses presented in this paper can be found on the `OSF` page for this study: `https://osf.io/up3rv/`.

Theory based method, H_{t_s}

For $H_{t_s} : \mu_{Distance} < \mu_{Intermediate} < \mu_{Closeness}$, the desired support was $BF = 3$ for either hypothesis. In the replication sample there were 44 and 38 participants in the

groups. The data eventually provided minimal support for the replication hypothesis, $BF_{t_sc} = 2.10$, but support does not reach the desired level. The difference between the means of the Distance group ($M = 5.31$, $SD = 1.15$), the Intermediate group ($M = 5.31$, $SD = 1.07$), and the mean of the Closeness group ($M = 5.44$, $SD = 0.83$) is not large enough (Cohen's $d = 0.13$ between both Distance and Closeness, and between Intermediate and Closeness), nor is the ordering convincingly out of order, to render the required evidence in favor of either the replication hypothesis or its complement.

Therefore, Bayesian updating (Rouder, 2014) can be applied here to obtain stronger support for either the replication hypothesis or its complement. Since we are not able to collect more data ourselves, we duplicate the dataset of the replication study. *This is not a recommended procedure!* It is used here only to demonstrate how updating works. We will assume that we have collected additional data with coincidentally the same sample means, standard deviations, and sample size as the first half. The BF_{t_sc} is calculated for the updated dataset. Now, the result is $BF_{t_sc} = 2.66$ (versus the previous $BF_{t_sc} = 2.10$). The extended dataset gives almost 3 times more support for the replication hypothesis than for the complement. Since the desired amount of support is still not achieved, it is reasonable to continue collecting data until the threshold of BF = 3 is achieved in favor of the replication hypothesis or its complement. *Note again that in a real situation the researcher would need to actually collect more data for Bayesian updating, which we did not do.*

After adding more data, in this case triplicating the original data, BF_{t_sc} is calculated again. Now, the result is $BF_{t_sc} = 3.10$ (versus the previous $BF_{t_sc} = 2.66$). The extended dataset gives more than three times as much support for the replication hypothesis over the complement. Now, the desired amount of support is achieved, we have a preference for H_{t_s}, however this preference comes with an error probability equal to $PMP_c = .24$.

Theory based method, $H_{t_{min_1}}$ and $_{t_{min_2}}$

For the minimal difference hypothesis two hypotheses were formulated: $H_{t_{min1}}$ ($\mu_{Distance} + 0.41 < \mu_{Intermediate} + 0.19 < \mu_{Closeness}$) and $H_{t_{min2}}$ ($\mu_{Distance} + 0.21 < \mu_{Intermediate} + 0.10 < \mu_{Closeness}$). Both hypotheses were formulated to show the impact of interpreting the minimal difference of 0.2 *SD*, either between adjacent groups or between the most extreme group means. A desired support of BF = 3 was set for this hypothesis. With approximately $n = 40$ per group we can see that the actual difference is not in agreement with the hypothesis. The mean of the Closeness group is 0.13 higher than both the mean of the Distance and the mean of the Intermediate group. Only the difference between the means of the Intermediate and the Distance group are according to the situation describe in $H_{t_{min1}}$. All other means are not in accordance with either hypothesis. The resulting BFs were: $BF_{t_{min1}c} = 0.57$ and $BF_{t_{min2}c} = 0.11$. There is no convincing support to prefer $H_{t_{min1}}$ or its complement: $PMP_{t_{min1}} = .36$ and $PMP_c = 0.64$, however, updating could be used to reach the desired level of support. There is clear support for H_c over $H_{t_{min}}$, $BF_{t_{min}c} = .11$, that is, $BF_{ct_{min}} = 9.10$ with $PMP_{t_{min}} = .10$ and $PMP_c = .90$.

Results based method, H_{r_M}

For the second method we consider the hypothesis H_{r_M} (H_{r_M} : $5.28 < \mu_{Closeness} < 5.94$ & $4.90 < \mu_{Intermediate} < 5.56$ & $4.53 < \mu_{Distance} < 5.19$). A desired support of BF = 10 was set for this hypothesis. The intervals appear small enough to formulate meaningful conclusions. Some overlap between adjacent means is present, but the extremes do not overlap. Furthermore, the interval cover around 10% of the scale, which leaves enough room for scores outside the intervals. We can continue with the results based method. With approximately $n = 40$ per group, two out of three means ($M_{Closeness} = 5.44$, $SD = 0.83$ & $M_{Intermediate} = 5.31$, $SD = 1.15$) from the replication study fall within their respective confidence intervals from the hypothesis. The mean of the Distance group ($M = 5.31, SD = 1.07$) does not fall within its hypothesized confidence interval. This results in some support for the replication hypothesis, $BF_{r_Mc} = 1.92$. There is some reason to further explore H_{r_M}, though there is no convincing support according to the standards set by us to prefer the presented hypothesis or its complement, the Bayesian error probability associated with a preference for H_{r_M} is $PMP_c = .34$. Updating could, in principle, be used to obtain stronger evidence for either hypothesis.

Results based method, H_{r_D}

Lastly the interval hypothesis specifying the difference between the means (H_{r_D} : $-0.09 < \mu_{Closeness} - \mu_{Intermediate} < 0.85$ & $-0.10 < \mu_{Intermediate} - \mu_{Distance} < 0.84$) is considered. For this hypothesis a desired support of BF = 10 was set. The difference scores are measured on a scale from -6 to 6. This means that the intervals cover about 10% of the scale, which leaves enough room for scores outside the intervals. We can continue with the results based method. Both intervals include the actual difference between the groups means found in the replication study, this results in $BF_{r_Dc} = 4.44$. There is support although not convincing enough for our goals to prefer either the replication hypothesis nor its complement. To achieve this level of support, updating could be used.

Conclusion

In reality, researchers would evaluate one replication hypothesis. That we evaluated four hypotheses was only to show the versatility of our approach. But it may be clear that most BFs show no clear preference for either hypothesis. The only BF that gives a clear answer shows strong support for the complement. All in all, we conclude that the results of Williams and Bargh (2008) are not convincingly corroborated by Joy-Gaba et al. (2012).

Example 2 - Janssen, Schirm, Mahon, and Caramazza

For the second example, the original article of Janssen et al. (2008) and the replication of Study 1 in this paper by Galak (2012) will be used. The decision was made to conduct the replication using the simple ordering hypothesis from the theory based method, because the goal of this replication is to test the general conflicting theories presented in the introduction of Janssen et al. (2008). This second example will illustrate a situation where competing theories are tested against each other, as well as a hypothesis specified using an equality constraint and an inequality constraint.

Step 1 - Reading the introduction section

Reading the introduction section

The goal of the study is to test whether the semantic interference effect arises in a delayed naming task. Participants were shown 20 pictures of common objects, half with low frequency names (used 1 to 9 times per million words) and half with high frequency names (used 72 to 724 times per million). In each trial, one of the 20 pictures (e.g., car) was shown on the screen. After 1000 ms, a distractor word was shown beneath the picture. The distractor could be either related (e.g., truck) or unrelated (e.g., table). Participants were instructed to name the picture as soon as the distractor word appeared on the screen. The semantic interference effect claims that a related distractor word makes it easier to find the correct word, so the response time is shorter when the distractor word is related. Since the participant sees the picture for some time, it is expected that the frequency (1-9 times or 72-724 times per million) has no effect on the time to give the proper response. If the semantic interference effect arises at a post-lexical level of processing, that is, after thinking about what to say, the distractor word would influence reaction time. The researchers wanted to investigate whether this effect appears in such a delayed naming task.

Step 2 - Coding of statements in the introduction section

From the introduction section a number of statements were coded that related to the hypotheses of interest to the researchers of the original paper:

- "The primary source of empirical evidence cited in support of lexical selection by competition is the semantic interference effect (...): Naming a picture of an object (e.g., CAR) is slower in the context of a semantic category coordinate distractor word (e.g., truck) compared to an unrelated distractor word (e.g., table)." (Janssen et al., 2008, p.2)
- "If the semantic interference effect arises at the level of lexical selection, then the semantic interference effect will not be observed when participants delay their picture naming responses." (Janssen et al., 2008, p.3)
- "However, if the semantic interference effect arises at a postlexical level of processing, then semantic interference should be observed in a delayed naming task." (Janssen et al., 2008, p.3)
- "If participants have already retrieved the lexical representations corresponding to the target picture names in the delayed naming condition at the time the cue is presented, then there should be no effect of the frequency of the target pictures on naming latencies." (Janssen et al., 2008, p.4)

Step 3 - Reading the methods & results section

The methods section makes clear that a 2 (Frequency: high; HF, vs low; LF) x 2 (Relatedness: related; rel, vs unrelated; unrel) ANOVA design was used. The dependent variable was the time the participants needed to name the object.

Step 4 - Coding of statements in the methods & results section

From the results section, the three statements with respect to hypothesis testing were coded. Note that the interest is on what is being tested, not on the results of the test:

- "The R[esponse] T[ime] analysis revealed a main effect of semantic relatedness, (...) ; $F_2(1, 38) = 4.4, MSE = 1,137.2, p < .05$, indicating slower RTs in the semantically related than the unrelated condition." (Janssen et al., 2008, p.6)

- Neither the main effect of frequency nor (...) reached significance (all $Fs < 1$). (Janssen et al., 2008, p.6)
- Neither (...) nor the interaction between frequency and semantic relatedness reached significance (all $Fs < 1$). (Janssen et al., 2008, p.6)

Step 5 - Selecting the statements from the introduction section that are actually tested

Based on the statements from the results section, it appeared that the first statement was not tested, it was merely a description of the effect under investigation in statements two and three. However, Statements 2 and 3 were tested.

Step 6 - Translating statements to informative hypotheses

Statements remaining from Step 5 will now be translated into simple ordering hypotheses:

- "If the semantic interference effect arises at the level of lexical selection, then the semantic interference effect will not be observed when participants delay their picture naming responses." can be represented as:
  $\mu_{1A} + \mu_{2A} = \mu_{1B} + \mu_{2B}$
- "However, if the semantic interference effect arises at a postlexical level of processing, then semantic interference should be observed in a delayed naming task." can be represented as:
  $\mu_{1A} + \mu_{2A} > \mu_{1B} + \mu_{2B}$
- "If participants have already retrieved the lexical representations corresponding to the target picture names in the delayed naming condition at the time the cue is presented, then there should be no effect of the frequency of the target pictures on naming latencies." can be represented as:
  $\mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B}$

The first two statements describe the same main effect, but with different outcomes. One suggests that there is a main effect: the reaction time is lower with unrelated distractors than with related distractors, the other describes the absence of this effect: both groups score equally on average. The third statement describes the other main effect, that of frequency. This results in the following replication hypotheses:

$$H_{t_1} : \mu_{1A} + \mu_{2A} = \mu_{1B} + \mu_{2B} \;\&\; \mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B},$$

$$H_{t_2} : \mu_{1A} + \mu_{2A} > \mu_{1B} + \mu_{2B} \;\&\; \mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B},$$

using the replication data, each can be tested against its complement, and, additionally, both can be tested against each other.

Determining the desired support

From the literature it is apparent that there is a clear and established theory regarding the effect of frequency on the reaction time. For the relatedness of distractors, there are contradicting ideas and theories. For this reason, the desired support is set to 5 for both of the replication hypotheses versus their respective complements. However, the replication hypotheses can also be tested against each other, and to gather some idea on which hypothesis needs further exploration, here a minimum BF of 5 is also chosen.

Calculate *BF* for the replication data

Three BFs are calculated: each replication hypotheses versus its complement and the replication hypotheses versus each other. Note that, for the same reason highlighted for interval hypotheses, here too `bain` was used with `fraction = 3` for all analyses. The statistics for both the original and the replication study are displayed in Table 3.

- $H_{t_1} : \mu_{1A} + \mu_{2A} = \mu_{1B} + \mu_{2B}$ & $\mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B}$ versus its complement. The result of the analysis, $BF_{t_1c} = 4.57$, indicates a slight preference for the first replication hypothesis, the error probability associated with a preference for H_{t_1} equals $PMP_c = .18$. This is corroborated by the η^2's for both main effects that can be found in Table 5, it is .02 for the first part of H_{t_1} and .03 for the second part, that is, both are rather small.
- $H_{t_2} : \mu_{1A} + \mu_{2A} > \mu_{1B} + \mu_{2B}$ & $\mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B}$ versus its complement. The result of this analysis, $BF_{t_2c} = 2.24$, lends hardly any support for the replication hypothesis versus its complement, the error probability associated with a preference for H_{t_2} equals $PMP_c = .31$.
- H_{t_1} versus H_{t_2}. When the two replication hypotheses are tested versus each other, the support for the first is hardly larger than for the second, $BF_{t_1t_2} = 2.04$, the error probability associated with a preference for H_{t_1} equals $PMP_{t_2} = .33$.

Table 3

Descriptive statistics and results of the original study of Janssen, Schirm, Mahon, & Caramazza (2008) and its replication by Galak (2012).

	Original study by Janssen, Schrim, Mahon, & Caramazza (2008)				Replication study by Galak (2012)			
	Group 1A	Group 1B	Group 2A	Group 2B	Group 1A	Group 1B	Group 2A	Group 2B
	1A	1B	2A	2B	1A	1B	2A	2B
M (*SD*)	nr[1]	nr	nr	nr	825.72 (41.51)	815.15 (41.63)	834.04 (53.45)	832.42 (34.46)
95% CI around *M*					(807.53, 843.91)	(796.90, 833.40)	(810.61, 857.47)	(817.32, 847.52)
n	nr	nr	nr	nr	20	20	20	20
$H_0 : \mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B}$	$p > .05$				$p = .27, \eta^2 = .03$			
$H_0 : \mu_{1A} + \mu_{2A} = \mu_{1B} + \mu_{2B}$	$p < .05, \eta^2 =$ nr				$p = .43, \eta^2 = .02$			
$H_0 : \mu_{1A} - \mu_{1B} = \mu_{2A} - \mu_{2B}$	$p > .05$				$p = .56, \eta^2 = .009$			
$H_{t_1} : \mu_{1A} + \mu_{2A} = \mu_{1B} + \mu_{2B}$ & $\mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B}$					$BF_{t_1 c} = 4.57$			
$H_{t_2} : \mu_{1A} + \mu_{2A} > \mu_{1B} + \mu_{2B}$ & $\mu_{1A} + \mu_{1B} = \mu_{2A} + \mu_{2B}$					$BF_{t_2 c} = 2.24$			
					$BF_{t_1 t_2} = 2.04$			

Note. [1] = This information is not reported (nr), nor can it be calculated based on the reported information. 2. The means are all measured in milliseconds.

Conclusion

We chose 5 as the desired support for both replication hypotheses versus their respective complements and versus each other. H_{t_1} ($\mu_{1A}+\mu_{2A} = \mu_{1B}+\mu_{2B}$ & $\mu_{1A}+\mu_{1B} = \mu_{2A}+\mu_{2B}$) received more than 4 times as much support as its complement. H_{t_1} receives about double the support as H_{t_2} ($\mu_{1A}+\mu_{2A} > \mu_{1B}+\mu_{2B}$ & $\mu_{1A}+\mu_{1B} = \mu_{2A}+\mu_{2B}$). Both of these BFs do not meet the desired support. There is some indication that the original results of Janssen et al. (2008) formulated in terms of H_{t_1} were corroborated by Galak (2012). However, since the sample sizes of the replication study were relatively small (20 persons per group), and the desired support for either hypothesis was not achieved, it is relevant to consider collecting additional data for the replication study and, subsequently, to update the Bayes factors.

Discussion

Inspired by the Reproducibility Project: Psychology (OSC, 2012; 2015) two refined methods were offered to plan and evaluate replication studies. In the theory based method the replication hypothesis was formulated based on the introduction of the original paper. This method would be executed when a theory elaborated in the introduction of the original study is evaluated using a replication study. Two kinds of informative hypotheses can be formulated here, a simple ordering hypothesis or a minimal difference hypothesis. In the results based method, the results of the original study are used to formulate the replication hypothesis. Here, an interval hypothesis around the means or around the difference between means is formulated.

We believe both methods enable replication studies that meet the needs and desires of replicating researchers. They can determine how much the results of a replication study favor or contradict the original, using the BF. Where necessary, Bayesian updating allows for the collection of data until the desired degree of certainty is obtained. Our methods allow to formulate multiple concurrent hypotheses as was illustrated in Example 2, and provide flexibility in formulating the replication hypothesis. Our approach is a remedy against the "null-ritual" (Gigerenzer, 2004) and explicitly addresses the importance of testing the correct hypothesis (which is not always done, Cho & Abe, 2013).

For the evaluation of both methods some subjective decisions had to be made. For both methods the desired support has been set to BF = 3 or 10. For the theory based method one has to decide whether the minimal difference applies to each pair of group or to the extreme groups. For the results based method `fraction = 3` has been chosen. For the results based method we used 95%-intervals, 90% or 99% intervals could also have been used. For the results based method one has to decide whether to focus on means or a difference between means. Within this article no investigation on the impact of the choices has been executed. The focus in this paper is on the hypotheses one formulates. It is essential to test the correct hypothesis, and the main message from this study is how to formulate that hypothesis. Therein still lies a subjective decision, replicating researchers should critically assess when a result corroborates the original study. In some instances the difference between adjacent means is important, sometimes a focus on the difference between the smallest and the largest means is most relevant (for the theory based method). And whereas in some circumstances hypotheses should be based on the means, in other one prefers the difference between means. Researchers should critically assess which situation applies to their situation before conducting the replication study. A sensitivity analysis could be helpful to understand the effect of these types of decisions. The remainder of the paper explains the further steps necessary to analyse the replication hypothesis. During that process, for each of the subjective decisions the mainstream options were followed.

A major strength of these methods is the possibility of updating (Rouder, 2014). Updating makes sure that studies end with convincing results. No estimates, assumptions, or guesses are necessary, unlike for power analyses. The replicating researchers only need to determine what the stopping point is, i.e. what the desired amount of support is.

The use of the BF brings one more feature to the table which is not yet discussed: information synthesis (e.g., Kuiper et al., 2013). Though not exclusive to the BF, synthesis is very easily conducted with the BF. With information synthesis, multiple BFs can be combined to create a meta view on the support for a certain hypothesis. It is necessary that the BFs under consideration are calculated under the same conditions, that is, with the same hypotheses. In the case of the theory based method, this leads to an interesting possibility. If the introduction section of the original paper is constructed before collecting the data for that paper, there are two datasets (the original and the replication) able to provide information on the replication hypothesis. By calculating the BF_{tc} twice, and multiplying the two BF_{tc}s a new overall BF_{tc} is generated, combining the information from both the original and the replicating study. For the results based method this is not possible, since the replication hypothesis is based on the results of the original paper. This leads to an interesting discussion for theory based method: can we assume that the hypotheses from the introduction section are not influenced by

the data from that very same paper? Ideally; this assumption should be met, but is that also true in reality?

The approaches presented in this paper can easily be executed using the `R` package `bain` (https://informative-hypotheses.sites.uu.nl/software/bain/, 2019). The approach presented in this paper is therefore well within reach of psychological researchers that want to execute a replication study in order to evaluate the results of an original study.

Author Contact

E-mail: H.J.Leplaa@uu.nl, Postal address: Utrecht University, Faculty of Social and Behavioral Sciences, Padualaan 14, PO Box 80140, 3508 TC Utrecht, The Netherlands.

Conflict of Interest and Funding

The last author is supported by a fellowship of the Netherlands Institute for Advanced Study for the Humanities and Social Sciences (NIAS-KNAW) and the Consortium on Individual Development (CID) which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003).

Author Contributions

HJL, CR, and HH were involved in the initial research design. HJL wrote and revised the article in collaboration with HH, CR was actively involved in the review and edit of the first draft. HJL conducted the simulation studies the analyses. All authors approved the final manuscript. The first author (HJL) was the main author, the last author (HH) was the main supervisor on this project, the second author (CR) was the daily supervisor during the first steps of this project.

Acknowledgments

The starting point for this paper was the bachelor thesis by Stiekema (2017). We would like to acknowledge the support of Fayette Klaassen and Qianrao Fu in the early stages of this paper: their feedback and explanations were of great value for this project. The data used in this paper comes from the Open Science Collaboration Reproducibility Project: Psychology.

Open Science Practices

This article earned the Open Data and Open Code badge for making the data and code openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, is published in the online supplement.

References

Anderson, S., Kelley, K., & Maxwell, S. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, *28*, 1547–1562. https://doi.org/10.1177/0956797617723724

Anderson, S., & Maxwell, S. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*, 1–12. https://doi.org/10.1037/met0000051

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of personality and social psychology*, *71*(2), 230–244. https://doi.org/10.1037/0022-3514.71.2.230

Berger, J., & Pericchi, L. (2004). Training samples in objective bayesian model selection. *The Annals of Statistics*, *32*, 841–869. https://doi.org/10.1214/009053604000000229

Boeije, H. (2010). *Analysis in qualitative research*. Sage.

Brandt, M., IJzerman, H., Dijksterhuis, A., Farach, F., Geller, J., Giner-Sorolla, R., Grange, J., Perugini, N., Spies, J., & Van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

Cho, H.-C., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research*, *66*. https://doi.org/10.1016/j.jbusres.2012.02.023

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 1664–1078. https://doi.org/10.3389/fpsyg.2014.00781

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind,

but whose mind? *PLoS ONE*, *7*(1). https://doi.org/10.1371/journal.pone.0029081

Erdfelder, E., Faul, F., & Buchner, A. (1996). Gpower: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*, 1–11. https://doi.org/10.3758/BF03203630

Etz, A., & Vandekerckhove, J. (2016). A bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, *11*(2). https://doi.org/10.1371/journal.pone.0149794

Field, S., Hoekstra, R., Brinkmann, L., & Van Ravenzwaaij, D. (2019). When and why to replicate: As easy as 1, 2, 3? *Collabra: Psychology*, *5*(46). https://doi.org/10.1525/collabra.218

Fu, Q. (2022). *Sample size determination for bayesian informative hypothesis testing.* [Doctoral dissertation]. Utrecht University [Utrecht University Repository. https://dspace.library.uu.nl/handle/1874/416118].

Galak, J. (2012). Replication of study 1 by janssen, schirm, mahon, & caramazza (2008, jep:lmc) [Retrieved from https://osf.io/uhpyr/].

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Gigerenzer, G. (2004). In *Blackwell handbook of judgment and decision making*. Blackwell Publishing. https://doi.org/10.1002/9780470752937.ch4

Glaser, B. (1978). *Theoretical sensitivity*. The Sociology Press.

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximate adjusted fractional bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*, 229–261. https://doi.org/10.1111/bmsp.12110

Harms, C. (2018). A bayes factor for replications of anova results. *The American Statistician*, *73*, 327–339. https://doi.org/10.1080/00031305.2018.1518787

Hawthorne, J. (2021). *Inductive logic* [Retrieved December 2, 2022]. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/spr2021/entries/logic-inductive/

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. New York: Chapman; Hall/CRC.

Hoijtink, H. (2022). Prior sensitivity of null hypothesis bayesian testing. *Psychological Methods*, *27*(5), 804–821. https://doi.org/10.1037/met0000292

Hoijtink, H., Gu, X., Mulder, J., & Rosseel, Y. (2019a). Bayesian evaluation of informative hypotheses for multiple populations. *British Journal of Mathematical and Statistical Psychology*, *72*, 219–243. https://doi.org/10.1111/bmsp.12145

Hoijtink, H., Gu, X., Mulder, J., & Rosseel, Y. (2019b). Computing bayes factors from data with missing values. *Psychological Methods*, *24*, 253–268. https://doi.org/10.1037/met0000187

Hoijtink, H., Mulder, J., Van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the bayes factor. *Psychological Methods*, *24*, 539–556. https://doi.org/10.1037/met0000201

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Psychology*, *66*, 81–92. https://doi.org/10.1016/j.jesp.2015.09.009

Informative hypotheses [computer software] [Retrieved from https://informative-hypotheses.sites.uu.nl/software/bain/]. (2019).

Janssen, N., Schirm, W., Mahon, B., & Caramazza, A. (2008). Semantic interference in a delayed naming task: Evidence for the response exclusion hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 249–256. https://doi.org/10.1037/0278-7393.34.1.249

Joy-Gaba, J., Clay, R., & Cleary, H. (2012). Replication of keeping one's distance: The influence of spatial distance cues on affect and evaluation by williams & bargh (2008, psychological science) [Retrieved from https://osf.io/vnsqg/].

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. https://doi.org/10.2307/2291091

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A bayesian approach. *Psychological Methods*, *10*, 477–493. https://doi.org/10.1037/1082-989X.10.4.477

Kuiper, R., Hoijtink, H., Buskens, V., & Raub, W. (2013). Combining statistical evidence from several studies: A method using bayesian updating and an example from research on trust problems in social and economic exchange. *Sociological Methods and Research*, *42*, 60–81. https://doi.org/10.1177/0049124112464867

Lakens, D., Scheel, A., & Isager, P. (2018). Equivalence testing for psychological research: A tutorial.

Advances in Methods and Practices in Psychological Science, *1*, 259–269. https://doi.org/10.1177/2515245918770963

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2018). Replication bayes factors from evidence updating. *Behavior Research Methods*, *51*, 2498–2508. https://doi.org/10.3758/s13428-018-1092-x

Marsman, M., Schönbrodt, F., Morey, R., Yao, Y., Gelman, A., & Wagenmakers, E.-J. (2017). A bayesian bird's eye view of 'replications of important results in social psychology'. *Royal Society Open Science*, *4*. https://doi.org/10.1098/rsos.160426

Morey, R. D., & Lakens, D. (n.d.). *Why most of psychology is statistically unfalsifiable* [[Retrieved November 22, 2022]. https://raw.githubusercontent.com/richarddmorey/psychology%5C_resolution/master/paper/response.pdf

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. https://doi.org/10.1016/j.jmp.2015.11.001

Mulder, J. (2014). Prior adjusted default bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics and Data Analysis*, *71*, 448–463. https://doi.org/10.1016/j.csda.2013.07.017

O'Hagan, A. (1995). Fractional bayes factors for model comparison (with discussion) [Retrieved June 10, 2021, from http://www.jstor.org/stable/2346088]. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 99–138.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*(6), 657–660. https://doi.org/10.1177/1745691612462588

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 6251. https://doi.org/10.1126/science.aac4716

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? a statistical view of replicability in psychological science. *Perspectives on Psychological Science*, *11*(4), 539–544. https://doi.org/10.1177/1745691616646366

Popper, K. R. (1963). Science as falsification. *Conjectures and refutations*, *1*, 33–39. https://curiousphilosophy.net/2023/09/is-sex-binary--a-reasoned-objection-to-rationality-rules-in-the-pursuit-of-truth/uploads/pdfs/Science%5C_as%5C_Falsification%5C_%5C_Karl%5C_R%5C_%5C_Popper.pdf

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

Rouder, J. (2014). Optional stopping: No problem for bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308. https://doi.org/10.3758/s13423-014-0595-4

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

Simonsohn, U. (2015). Small telescopes detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Stiekema, J. (2017). *Bayesiaanse evaluatie van informatieve hypotheses als methode voor replicatiebeoordeling* [Unpublished bachelor's thesis].

Van Aert, R., & Van Assen, M. (2017). Bayesian evaluation of effect size after replicating an original study. *PLoS ONE*, *12*. https://doi.org/10.1371/journal.pone.0175302

Van Lissa, C., Gu, X., Mulder, J., Rosseel, Y., Van Zundert, C., & Hoijtink, H. (2020). Teacher's corner: Evaluating informative hypotheses using the bayes factor in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 292–301. https://doi.org/10.1080/10705511.2020.1745644

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457. https://doi.org/10.1037/a0036731

Williams, L., & Bargh, J. (2008). Keeping one's distance: The influence of spatial distance cues on affect and evaluation. *Psychological Science*, *19*, 302–308. https://doi.org/10.1111/j.1467-9280.2008.02084.x

Wilson, A. (2016). Exact replications in an inexact context: Commentary on ebersole et al. *Journal of Experimental Social Psychology*, *67*, 84–85. https://doi.org/10.1016/j.jesp.2015.12.008

Zondervan-Zwijnenburg, M., van de Schoot, R., & Hoijtink, H. (2019). Testing anova replications by

means of the prior predictive p-value [Retrieved on January 28, 2019]. https://doi.org/10.31234/osf.io/6myqh

Appendix

The Approximate Adjusted Fractional Bayes Factor

The Approximate Adjusted Fractional Bayes Factor (AAFBF) is implemented in the `R` package `bain` and was developed such that it can be applied in a wide range of statistical models for the evaluation of informative hypotheses in addition to the traditional null and alternative hypotheses. The statistical background of the AAFBF can be found in (Gu et al., 2018; Hoijtink, 2022; Hoijtink, Gu, et al., 2019a, 2019b), more practically oriented tutorials about the use of the AAFBF can be found in (Hoijtink, Gu, et al., 2019a; Van Lissa et al., 2020). In this appendix the AAFBF, simplified to the form it attains in the context of ANOVA models, will be presented.

The Bayes factor of H_i, which denotes either a null (H_0) or an informative hypothesis (H_i) versus the unconstrained hypothesis H_u can be written as the ratio of two marginal likelihoods:

$$BF_{iu} = \frac{m(D|H_i)}{m(D|H_u)} = \frac{f_i}{c_i}, \tag{A1}$$

where D denotes the data, that is, for an ANOVA, group membership and score on the dependent variable for each person. As was shown by Klugkist et al. (2005), this ratio of two marginal likelihoods can be rewritten in term of the ratio of the fit f_i and complexity c_i of H_i.

The "approximate" in AAFBF results from the fact that the fit is computed with respect to a normal approximation of the posterior distribution of $\mu = [\mu_1, ..., \mu_G]$, where G denotes the number of groups in the ANOVA:

$$g(\mu|D) = \mathcal{N}(\mu|\hat{\mu}, \Sigma), \tag{A2}$$

where $\hat{\mu}$ denotes estimates of the means, and Σ the covariance matrix of the estimates which is diagonal with elements σ^2/N_g for $g = 1, ..., G$, where σ^2 denotes the pooled within groups variance, and N_g the sample size in group G. Already for relatively small sample size per group (e.g., 20) the normal approximation becomes virtually indistinguishable from the true (a t-distribution) posterior distribution of the means. The fit is

$$f_i = \int_{\mu \in H_i} g(\mu|D) d\mu. \tag{A3}$$

It is illustrative to note that for hypotheses constructed using only inequality constraints, the fit is the proportion of the posterior distribution in agreement with H_i. In general it holds that the larger the fit the larger the support in the data for H_i.

The complexity is based on the "adjusted" "fractional" prior distribution corresponding the posterior distribution:

$$h(\mu|D) = \mathcal{N}(\mu|\mu_{adj}, \Sigma_{frac}), \tag{A4}$$

where the adjusted prior mean μ_{adj} is a value on the boundary of H_i and its complement H_c. For example, the complement of $H_0 : \mu_1 - \mu_2 = 0$ is H_c : "not H_i" and therefore any $\mu_{1,adj} = \mu_{2,adj}$ can be used as the prior mean. Analogously, a value on the boundary of $\mu_1 > \mu_2 > \mu_3$ and its complement would be $\mu_{1,adj} = \mu_{2,adj} = \mu_{13,adj}$. For interval hypotheses like $2 < \mu < 4$, μ_{adj} is set equal to the mean of the interval, that is, for setting the mean of the prior distribution this hypothesis is treated analogously as $\mu = 3$. The prior covariance matrix is based on a fraction of the information in the data with respect to the μ's, that is, $\Sigma_{frac} = \Sigma * 1/b$, where $b = [b_1, ..., b_G]$ with $b_g = 1/G * J/N_g$, denotes the fraction of information in group G used to specify the prior variance of the mean in group g. The parameter J will be elaborated in the next paragraph. The complexity is

$$c_i = \int_{\mu \in H_i} h(\mu|D) d\mu. \tag{A5}$$

It is illustrative to note that for hypotheses constructed using only inequality constraints, the complexity is the proportion of the prior distribution in agreement with H_i. This implies that the Bayes factor of an informative hypothesis versus its complement can be written as:

$$BF_{ic} = \frac{f_i}{c_i} / \frac{1 - f_i}{1 - c_i}. \tag{A6}$$

In general it holds that the larger the complexity the less specific, that is, the less parsimonious H_i.

The prior means are chosen in the manner elaborated in the previous paragraph to ensure that both BF_{0u} and BF_{ic} are consistent (Gu et al., 2018; Hoijtink, Gu, et al., 2019a), that is, if the sample size increases these Bayes factors go to 0 if H_u or H_c is the best hypothesis and to infinity if H_0 or H_i is the best hypothesis. This holds for any value of J. The prior variances are chosen in agreement with the minimal training sample principle (see, for example, Berger & Pericchi, 2004; O'Hagan, 1995), that is, what is the smallest sample size that allows estimation of the parameters of the ANOVA model (the answer is $G + 1$). In the AAFBF J is chosen not equal to the minimal training sample size, but equal to the number of independent constraints used to specify the hypothesis of interest. In case of $H_0 : \mu_1 = \mu_2 = \mu_3$ the number of independent constraints equals two, the same holds for $H_i : \mu_1 > \mu_2 > \mu_3$. Finally note, that the choice of J is only relevant if the goal is to evaluate a hypothesis specified using equality or inequality constraints, in all other cases (hypotheses specified using only inequality constraints that are not interval hypotheses) the Bayes factor is completely insensitive to the specification of J (Hoijtink, Mulder, et al., 2019; Mulder, 2014).