Research Master's programmes
Methodology & Statistics for the Behavioral, Biomedical, and Social Sciences (MSBBSS) *and* Educational Sciences: Learning in Interaction (EdSci)
Utrecht University, the Netherlands

MSc Thesis *Millitza Damaris Kroonenberg (3995356)*
TITLE: Replication of Cohen's d: Quantifying Evidence using the Bayes Factor
May 2016

Supervisor:
Prof. dr. Herbert Hoijtink

Preferred journal of publication: Psychological Methods
Word count: 5,669

Replication of Cohen's d: Quantifying Evidence using the Bayes Factor

Millitza D. Kroonenberg
Utrecht University

Abstract

This study introduces and demonstrates two new methods to quantify the evidence for replication of Cohen's d using the Bayes factor. The present study defines successful replication as the degree to which Cohen's d in the replication study is in line with the Cohen's d found in the original study. Non-successful replication is formalized in two hypotheses related to *relevant* and *nonzero* differences in Cohen's d respectively between the original and replication study. The Bayes factor that quantifies the relative evidence for successful replication versus relevant differences outperforms the Bayes factor that concerns nonzero differences in Cohen's d. The main conclusion of this paper is that researchers should include approximately 100 participants per group in both the original and replication study to be able to acquire convincing evidence in favor of or against successful replication, if differences of 0.50 in Cohen's d between the original and replication data set are deemed relevant. However, the required sample sizes increase up to 600 participants per group if differences in Cohen's d of 0.20 are considered to be relevant. The accompanying software quantifies the evidence for replication of Cohen's d for *t*-tests or post-hoc tests in analysis of variance (ANOVA), as is demonstrated using a data example.

Keywords: ANOVA, Bayes factor, Cohen's d, replication, *t*-test.

Interested readers may contact the author via millitza@gmail.com or Herbert Hoijtink via h.hoijtink@uu.nl for more information on the manuscript and accompanying `R` function. The `R` function and `Shiny` application are available via `http://www.github.com/millitza`. Address: Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, The Netherlands.

Introduction

Successful replication can increase the perceived and actual value and reliability of scientific studies (Cumming, 2008; Simons, 2014; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Replication of scientific studies is needed, because a single-study-based estimate of Cohen's d is often only an indication of a population effect size (Cumming, 2008). The call for replication may stimulate replication initiatives, but raises two important questions: 1. What is *successful* replication? and 2. What quantification of replication evidence is appropriate? These two questions should be treated consecutively, as the definition of successful replication determines the appropriate quantification of replication evidence (Anderson & Maxwell, 2016; Bayarri & Mayoral, 2002; Simonsohn, 2015).

To quantify the evidence in favor of or against successful replication at least two data sets are needed: one collected during an original study and one collected during its replication study. The original and replication studies both yield estimates of the effect sizes δ_{orig} and δ_{repl} respectively. The most commonly used effect size when investigating whether two group means differ is the standardized population mean difference or Cohen's d (Cumming & Finch, 2001; Cohen, 1992). This effect size is defined as $\delta_r = (\mu_1 - \mu_2)/\sigma$, for $r = orig, repl$. Here μ_1 and μ_2 are two population means and σ is the common intra-population standard deviation. Please note that μ_1, μ_2, and σ as well as data and sample sizes should have subscript r to denote whether they apply to the original or replication study. To improve readability, the subscript is not presented for these quantities. It will be clear from the context in which these quantities are presented to which study they apply.

Anderson & Maxwell (2016) recently identified six main goals for replication research. This section elaborates on four of these goals in the context of group mean comparison, before focusing on the two goals that relate to the notion of successful replication as adopted within this paper: replication studies in which Cohen's d is in line with the Cohen's d that resulted from the original study are considered successful replications, while non-corroborating replication results indicate non-successful replication.

One goal for replication research may be to study if the conclusions (Bayarri & Mayoral, 2002) or significance and direction of an effect (Anderson & Maxwell, 2016) obtained in the original and replication study are similar. The reproducibility project of the Open Science Collaboration (2015) compared the significance and direction of the original and replication effects as the main approach to assess successful replication. Replication studies that yielded estimates of Cohen's d that were significant and directionally in agreement with the original studies' estimates of Cohen's d, were identified as successful replications. Although the resulting dichotomous conclusions are clearcut, a second goal for replication studies may be appropriate if a researcher questions the outcomes of an original study. He or she may set up a replication study to infer a null effect (Anderson & Maxwell, 2016; Simonsohn, 2015), that is to study if δ_{repl} is more similar to δ_{orig}, or to 0. This goal drives research by Verhagen & Wagenmakers (2014) and Simonsohn (2015). A third goal may be to use the original study to determine the sample size in the replication study with the goal to analyze the replication dataset independently of the original data set. Here, the data from the original and replication study are not combined in the analyses in any way. In other replication studies the goal may be to quantify a joint population effect based on the results of multiple studies (e.g. meta-analysis), which is the fourth goal that Anderson

& Maxwell (2016) identified. The final two goals concern the interpretation of successful replication as used in this paper. According to Anderson & Maxwell (2016) researchers may study to what extent the results in a replication study are inconsistent (goal 5) or consistent (goal 6) with the results of the original study. The main rationale behind this approach is that there is uncertainty in effect size estimates of the original and replication study. This uncertainty should be taken into account when deciding if δ_{repl} is in line with δ_{orig}. In our view these two goals are strongly related and should be studied in conjunction. Therefore, the remainder of this section deals with methods that quantify replication evidence concerning these two goals for replication research.

Multiple researchers introduced methods that evaluate whether the standardized mean differences of the replication (δ_{repl}) and original study (δ_{orig}) are consistent or inconsistent. Anderson & Maxwell (2016) and Bayarri & Mayoral (2002) study the difference in effect sizes $\Delta\delta = \delta_{repl} - \delta_{orig}$. Maxwell, Lau, & Howard (2015) and Anderson & Maxwell (2016) propose the construction of a confidence interval for $\Delta\delta$ and the specification of a zone of equivalence around zero. In line with the sixth goal, concerning whether Cohen's d of the original and replication study are in line, they test if the confidence interval (almost) fully overlaps with this zone of equivalence. If this is the case, it is concluded that the replication attempt was successful. The fifth goal for replication research is studied by Anderson & Maxwell (2016) using the same confidence interval around $\Delta\delta$. If the confidence interval does not include zero there is evidence for non-successful replication. It must be noted that both approaches result in a dichotomous decision of successful or non-successful replication and therefore do not quantify the evidence for replication.

A second approach was introduced by Bayarri & Mayoral (2002). They quantify the relative evidence for $H_1 : \Delta\delta = 0$ and $H_2 : \Delta\delta \neq 0$ in a Bayes factor. This Bayes factor allows for a continuous quantification of evidence. However, one may criticize the nil-hypothesis $H_1 : \Delta\delta = 0$ to be too strict (Cohen, 1990). Additionally, H_2 must be formalized in a prior distribution for $\Delta\delta$. The precise specification of this prior distribution is important and difficult. However, Bayarri & Mayoral (2002) do not provide guidelines to specify the prior distribution for this hypothesis.

A third approach to assess whether the results of a replication study corroborate the results of the original study was used by Etz & Vandekerckhove (2016) to reanalyze the data of the Open Science Collaboration (2015). In this approach the original and replication data set are analyzed independently to obtain Bayes factors that quantify the relative evidence for $H_1 : \delta_r = 0$ and $H_2 : \delta_r \sim N(0, 1)$, for $r = orig, repl$. In the next step the resulting Bayes factors of the original and replication data are compared qualitatively. An advantage of this approach is that these analyses are readily available for many models. However, as is discussed by Etz & Vandekerckhove (2016), the values of the Bayes factors depend on both Cohen's d and sample size. Therefore, it does not provide a solid answer to the question if the results of a replication study are in line with the results of the original study if the sample sizes differ.

Our approach differs from the existing approaches, because it does not result in dichotomous conclusions. Rather, our approach quantifies the evidence for hypotheses relating to successful and non-successful replication. These hypotheses, in line with Mulaik, Raju, & Harhman (2016), "ought to test ... that the effect is equal to the value estimated in the previous study, which one judged to be significantly different from a zero effect." (p.91). To achieve this, our approach makes use of two Bayes factors, each quantifying the evidence in favor and against successful replication, that are based on data from the original and replication study. The original data set provides information for the explicit formalization of the hypotheses regarding successful and non-successful replication.

The current paper introduces and demonstrates this new way to quantify evidence for effect size replication using the Bayes factor in the framework of comparing two group means. The following section will discuss the formalized hypotheses, density of the data, prior distributions, and Bayes factor. Subsequently, we will present the resulting Bayes factors for data sets that differ in sample sizes and difference in Cohen's d between the original and replication study. The penultimate section presents the application of the developed Bayes factors for the reanalysis of a data example from the Open Science Framework. The final section summarizes and discusses the introduced approaches. Researchers who are interested to quantify the evidence for replication of Cohen's d for their own data are invited to download and apply the accompanying `Shiny` application or `R` code (`http://www.github.com/millitza`).

Quantification of replication evidence for Cohen's d

Evidence for replication of Cohen's d can be quantified in Bayes factors that weigh the relative evidence for successful and non-successful replication. This section introduces one hypothesis for successful replication (H_1) and two hypotheses that reflect non-successful replication (H_2 and H_3). In this paper consistency of δ_{repl} with δ_{orig} is considered successful replication, while inconsistency reflects non-successful replication. Analysis of the replication data under these three hypotheses result in two separate Bayes factors, BF_{12} and BF_{13}, that quantify the relative support in the replication data for the competing hypotheses. This section formalizes the competing hypotheses in prior distributions on δ_{repl}, discusses the density of the data, and elaborates on the calculation and interpretation of the Bayes factors.

Hypotheses and prior distributions

The three competing hypotheses are displayed in Figure 1. Hypothesis 1 (H_1): δ_{repl} is consistent with δ_{orig}, is represented by the solid curve in Figure 1. This hypothesis reflects successful replication and expresses exactly the information on the effect size δ that was obtained in the original study. This hypothesis can be expressed by choosing the prior on δ_{repl} to equal the posterior distribution of the effect size in the original study:

$$\pi(\delta_{repl}|H_1) = N(\bar{\delta}_{orig}, \gamma^2) \ , \tag{1}$$

where $\bar{\delta}_{orig}$ and γ^2 are the posterior mean and variance of δ_{orig}. Appendix A discusses how these estimates are obtained. Like the following two hypotheses, this hypothesis concerns only the effect size parameter δ_{repl} and does not constrain the distributions of μ_1 and σ^2. Therefore, the prior distributions for μ_1 and σ^2 are non-informative and independent of the hypotheses (see Appendix B). However, the prior distribution on δ_{repl} differs per hypothesis and constitutes a population of all values that this parameter can take on under the specific hypothesis (Gelman, Carlin, Stern, & Rubin, 2004, p. 39).

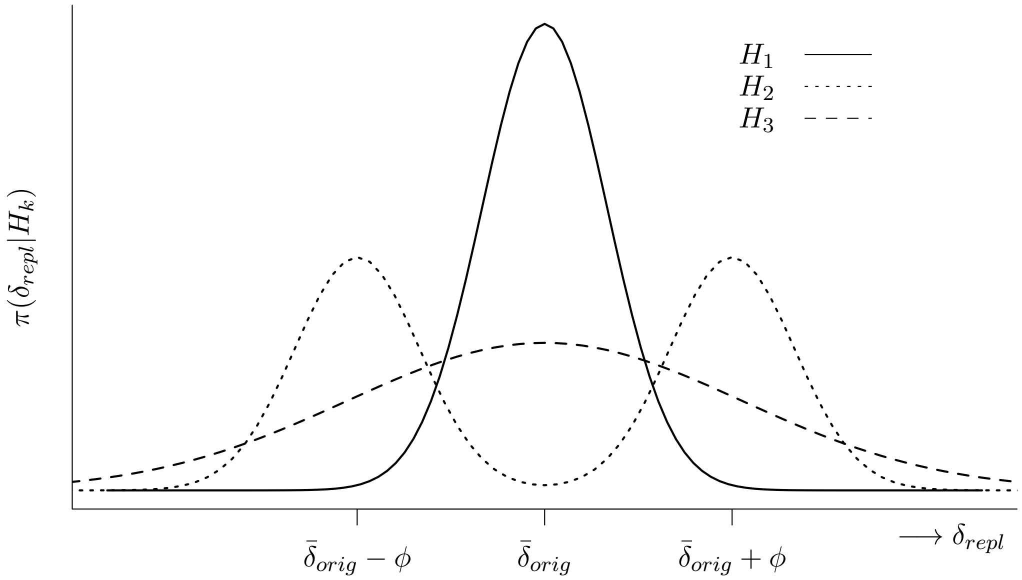


Figure 1. Example of three prior distributions for competing hypotheses H_k for $k = 1, 2, 3$. H_1 summarizes the results of the original study, while H_2 represents the competing hypothesis that the replication results deviate an amount of ϕ from the original results. The final hypothesis, H_3, reflects nonzero differences from $\bar{\delta}_{orig}$.

The dotted curve in Figure 1 represents Hypothesis 2 (H_2): δ_{repl} is relevantly different from δ_{orig}. In some situations researchers may not be interested in small differences in Cohen's d between studies. For example, the clinical or practical relevance of $\delta_{repl} = 0.30$ may be similar to $\delta_{orig} = 0.40$. Therefore, the researcher should decide what difference in Cohen's d is relevant and construct a prior distribution accordingly. The current study formalizes this hypothesis as follows:

$$\pi(\delta_{repl}|H_2) = \frac{1}{2}\pi_{min} + \frac{1}{2}\pi_{max} = \frac{1}{2}N(\bar{\delta}_{orig} - \phi, \gamma^2) + \frac{1}{2}N(\bar{\delta}_{orig} + \phi, \gamma^2) \ , \tag{2}$$

where ϕ represents the size of relevant deviation from $\bar{\delta}_{orig}$. This distribution is a mixture of two normal distributions with variance γ^2. The variance around each of the two maxima is equal to the uncertainty about δ in the original study. However, the location of the maxima deviates from the maximum of the prior distribution under H_1 by an amount of ϕ. For example, if $\bar{\delta}_{orig} = 0.30$ and the effect is deemed relevant if $\Delta\delta = 0.20$ (thus $\phi = 0.20$), the maxima should be located at $(0.10, 0.50)$. When H_1 competes with H_2 the question effectively is whether δ_{repl} is more in line with δ_{orig} or with $\delta_{orig} \pm \phi$.

The appropriate value for ϕ should be determined by the researcher that conducts a replication study. Researchers should assess the impact of differing effect sizes of the study at hand in their research field. The deviation ϕ can be chosen such that it reflects the smallest difference in Cohen's d that would be deemed relevant.

The third hypothesis, reflecting non-successful replication, is represented by the dashed curve in Figure 1. Hypothesis 3 (H_3): δ_{repl} deviates from δ_{orig} by a nonzero difference is represented as follows:

$$\pi(\delta_{repl}|H_3) = N(\bar{\delta}_{orig}, \gamma^2 + \phi^2) \ . \tag{3}$$

The maximum of this normal distribution is $\bar{\delta}_{orig}$ and is identical to the maximum of $\pi(\delta_{repl}|H_1)$. The variance of $\pi(\delta_{repl}|H_3)$ equals the total variance of $\pi(\delta_{repl}|H_2)$: $\gamma^2 + \phi^2$ (Behboodian, 1970). Thereby this prior distribution is less informative than $\pi(\delta_{repl}|H_1)$. Note that for both H_2 and H_3 the prior variance of δ_{repl} depends on γ^2, resulting from the original study, and ϕ, the researcher specified minimally relevant difference.

Density of the data

For a two-group analysis of variance with sample sizes n_1 and n_2, corresponding group means, μ_1 and μ_2, and within population variance σ^2 the density of the data is:

$$f(\boldsymbol{y}|\boldsymbol{D}, \mu_1, \delta_r, \sigma^2) = \prod_{i=1}^{N=n_1+n_2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \mu_1 + D_i\delta_r\sigma)^2} \qquad \text{for } r = orig, repl \ , \tag{4}$$

where $\boldsymbol{y} = [y_1, y_2, \ldots, y_i, \ldots, y_N]$, $\boldsymbol{D} = [D_1, D_2, \ldots, D_i, \ldots, D_N]$, and $D_i = 0$, for participants in Group 1 and $D_i = 1$ for participants in Group 2. Using $D_i\delta_r\sigma = D_i(\frac{\mu_1 - \mu_2}{\sigma})\sigma = D_i(\mu_1 - \mu_2)$ it can easily be verified that the group mean equals μ_1 for participants in Group 1 ($D_i = 0$) and μ_2 for participants in Group 2 ($D_i = 1$). This parameterization of the density of the data allows the use of an informative prior on the effect size δ_r.

Bayes factor

The density of the data and prior distributions are needed to calculate the Bayes factor. The Bayes factor (see, for example, Kass & Raftery, 1995) quantifies the relative evidence for two competing hypotheses. It is defined as the ratio of the marginal likelihoods of the data under the two competing hypotheses:

$$\begin{aligned}\mathrm{BF}_{1k} &= \frac{m(\boldsymbol{y}|\boldsymbol{D}, H_1)}{m(\boldsymbol{y}|\boldsymbol{D}, H_k)} \\ &= \frac{\iiint f(\boldsymbol{y}|\boldsymbol{D}, \mu_1, \delta_{repl}, \sigma^2)\pi(\mu_1, \delta_{repl}, \sigma^2|H_1)\ \mathrm{d}\mu_1\ \mathrm{d}\delta_{repl}\ \mathrm{d}\sigma^2}{\iiint f(\boldsymbol{y}|\boldsymbol{D}, \mu_1, \delta_{repl}, \sigma^2)\pi(\mu_1, \delta_{repl}, \sigma^2|H_k)\ \mathrm{d}\mu_1\ \mathrm{d}\delta_{repl}\ \mathrm{d}\sigma^2} \qquad \text{for } k = 2, 3 \ ,\end{aligned} \tag{5}$$

where $\pi(\mu_1, \delta_{repl}, \sigma^2|H_k) = \pi(\mu_1)\pi(\delta_{repl}|H_k)\pi(\sigma^2)$, see Appendix B. Values of Bayes factors that are greater than one, express stronger evidence in favor of H_1 than for H_k. This can be interpreted as successful replication. On the other hand, if the value of a Bayes factor is smaller than 1, the evidence for H_k is stronger than for H_1, reflecting non-successful replication. For example, if $\mathrm{BF}_{12} = 3$, the evidence for successful replication is three times stronger than for non-successful replication as expressed by H_2. The more the Bayes factor deviates from 1, the larger the relative evidence in favor of one of the competing hypotheses. The continuous scale of the Bayes factor does not result in dichotomous decisions, but provides a quantification of relative evidence for the competing hypotheses.

The challenge in calculating the Bayes factor is to acquire a reliable estimate of the marginal likelihood of the data under both hypotheses. Chib (1995) and Chib & Jeliazkov

(2001) proposed a stable and efficient method to estimate the natural logarithm of the marginal likelihoods of the replication data. Applied to our situation this results in:

$$\begin{aligned}\ln m(\boldsymbol{y}|H_k) = {} & \ln f(\boldsymbol{y}|\boldsymbol{D}, \mu_{1*}, \delta_{repl*}, \sigma_*^2) + \ln \pi(\mu_{1*}) + \ln \pi(\delta_{repl*}|H_k) + \ln \pi(\sigma_*^2) \\ & - \ln p(\mu_{1*}, \delta_{repl*}|\boldsymbol{y}, \boldsymbol{D}, \sigma_*^2) - \ln p(\sigma_*^2|\boldsymbol{y}, \boldsymbol{D}) \qquad \text{for } k = 1, 2, 3 \ . \end{aligned} \tag{6}$$

This expression can be evaluated at any set of assessment values of the parameters, $\mu_{1*}, \delta_{repl*}, \sigma_*^2$. In this paper the means of the posterior distributions of the parameters were used as the assessment values (see Appendix B). The density of $f(\boldsymbol{y}|\boldsymbol{D}, \mu_{1*}, \delta_{repl*}, \sigma_*^2)$ can be evaluated using Equation 4. The prior distributions $\pi(\mu_1)$ and $\pi(\sigma^2)$ can be found in Appendix B. For each hypothesis, $\pi(\delta_{repl*}|H_k)$ can be computed using Equations 1-3. Appendix C elaborates how the fifth and sixth element of Equation 6 can be evaluated.

Evidence for replication of Cohen's d

This section presents Bayes factors that quantify the evidence for replication of Cohen's d for multiple generated data sets. The results show the general properties of BF_{12} and BF_{13} for original and replication data sets with various sample sizes. Additionally, the influence of various differences in Cohen's d between the original study and replication study ($\Delta\delta$) is shown for four values of ϕ.

Data and conditions

Original and replication data sets with equal sample sizes for the two groups ($n = 20, 40, 100$) were generated using BIEMS (Mulder, Hoijtink, & Leeuw, 2012). BIEMS can generate data sets such that Cohen's d in the sample and population are equal. Cohen's d in all original data sets was exactly equal to 0.00. In the replication data sets Cohen's d was equal to 0.00, 0.20, 0.50, or 0.80, reflecting null, small, medium, and large effects according to Cohen (1992). As a result $\Delta\delta = 0.00, 0.20, 0.50, 0.80$.

Analysis

Bayes factors were computed for each pairwise combination of original data sets ($\delta_{orig} = 0.00; n_{orig} = 20, 40, 100$) and replication data sets ($\delta_{repl} = 0.00, 0.20, 0.50, 0.80; n_{repl} = 20, 40, 100$) for $\phi = 0.20, 0.35, 0.50, 0.80$. This choice reflects a small, small-medium, medium or large relevant difference in Cohen's d between the original and replication study. The Bayes factors were interpreted using guidelines on the strength of evidence as proposed by Kass & Raftery (1995). In our interpretation of Bayes factors, we will use the guideline that $\text{BF}_{1k} > 3$ indicates positive evidence in favor of H_1, while $\text{BF}_{1k} < \frac{1}{3}$ indicates positive evidence in favor of H_k.

A function and `Shiny` (Chang, Cheng, Allaire, Xie, & McPherson, 2016) application were developed in `R` (Version 3.2.2; R Core Team, 2015) to compute BF_{12} and BF_{13}. Figure 2 shows an overview of how the original and replication data set are used in calculating the Bayes factors. As can be seen from Figure 2 first $\bar{\delta}_{orig}$ and γ^2 are estimated using the original data set and non-informative prior distributions on μ_1, δ_{orig}, and σ^2 (see Appendix A). As elaborated in the previous section, $\bar{\delta}_{orig}$, γ^2, and ϕ are used to construct $\pi(\delta_{repl}|H_k)$

$$
\begin{array}{llcll}
 & & & \pi(\mu_1) = N(0, 10000) & & \\
\textbf{Step 1.} & \text{Original data} & + & \pi(\delta_{orig}) = N(0, 10000) & \longrightarrow & \bar{\delta}_{orig}, \gamma^2 \\
 & & & \pi(\sigma^2) = \Gamma^{-1}(0.001, 0.001) & & \\
 & & & & & \\
\textbf{Step 2.} & \bar{\delta}_{orig} \quad + \quad \gamma^2 & + & \phi & \longrightarrow & \pi(\delta_{repl}|H_k) \\
 & & & & & \\
 & & & \pi(\mu_1) = N(0, 10000) & & \\
\textbf{Step 3.} & \text{Replication data} & + & \pi(\delta_{repl}|H_k) & \longrightarrow & \text{BF}_{12}, \text{BF}_{13} \\
 & & & \pi(\sigma^2) = \Gamma^{-1}(0.001, 0.001) & &
\end{array}
$$

Figure 2. Overview of the required steps to calculate BF_{12} and BF_{13}.

for $k = 1, 2, 3$ in the second step. In the third step, the prior distributions on δ_{repl}, non-informative priors on μ_1 and σ^2 and the replication data are used to calculate BF_{12} and BF_{13} (see Appendix B and C).

As described in Appendices A and B, sampling from the posterior distributions of μ_1, δ_r and σ^2 given the original or replication data set is an iterative process. For H_1 and H_3 60,000 samples were drawn from the joint posterior distribution including 10,000 burn-in iterations for the Metropolis-Hastings-within-Gibbs sampler. Under H_2 120,000 samples were drawn: 60,000 for each of the normal distributions that were used to specify $\pi(\delta_{repl}|H_2)$. For both sequences a burn-in period of 10,000 iterations was used.

Results

Table 1 shows the evidence for successful replication of Cohen's d relative to H_2 and H_3 for all conditions. Two conclusions can be drawn from the calculated Bayes factors for the generated data sets. First, it is not possible to acquire convincing evidence in favor of or against successful replication if the original sample consists of 20 participants per group, unless $\phi \geq 0.80$. The Bayes factors in Table 1a-c for $n_{orig} = 20$ do not exceed the value 3.10 if $\Delta\delta = 0.00$. Table 1d shows that BF_{12} increases rapidly to values larger than 3 if $\phi = 0.80$. However, since the goal is to obtain evidence for replication a value of ϕ of 0.80 seems too extreme in most situations. Many researchers interpret Cohen's d = 0.80 as a large effect. If one study would yield $\delta = 0.00$ and the other study yields $\delta = 0.80$ this would not be considered the minimum relevant difference in Cohen's d. The underlying reason for indecisive evidence for small samples and $\phi \leq 0.50$, is that γ^2 is very large for small sample sizes, which leads to very similar prior distributions on δ_{repl} for H_1, H_2, and H_3. As a result, the replication data can hardly differentiate between successful or non-successful replication. It may sound obvious that studies with such small sample sizes cannot yield evidence in favor of or against replication, but replication studies often include too few participants. In the data example that is discussed in this paper the original study included only 16 participants per group, while the replication study sample size was 24 participants per group. These underpowered studies probably result from flawed sample

Table 1

BF_{12} and BF_{13} for Original and Replication Data Sets with n = 20, 40, 100, and a Difference in Effect Size $\Delta\delta$ = 0.00, 0.20, 0.50, 0.80.

a) $\phi = 0.20$

	$\Delta\delta$	0.00		0.20		0.50		0.80	
n_{orig}	n_{repl}	BF_{12}	BF_{13}	BF_{12}	BF_{13}	BF_{12}	BF_{13}	BF_{12}	BF_{13}
20	20	1.15	1.10	1.12	1.08	1.00	1.00	0.85	0.87
	40	1.17	1.13	1.13	1.10	0.95	0.95	0.72	0.74
	100	1.20	1.15	1.14	1.11	0.88	0.90	0.60	0.61
40	20	1.22	1.13	1.18	1.10	0.98	0.96	0.75	0.76
	40	1.31	1.18	1.20	1.12	0.84	0.84	0.51	0.51
	100	1.40	1.25	1.19	1.13	0.63	0.67	0.29	0.25
100	20	1.33	1.15	1.26	1.11	0.96	0.92	0.66	0.64
	40	1.56	1.25	1.32	1.13	0.70	0.68	0.33	0.27
	100	1.95	1.42	1.25	1.10	0.32	0.31	0.08	0.03

b) $\phi = 0.35$

	$\Delta\delta$	0.00		0.20		0.50		0.80	
n_{orig}	n_{repl}	BF_{12}	BF_{13}	BF_{12}	BF_{13}	BF_{12}	BF_{13}	BF_{12}	BF_{13}
20	20	1.51	1.28	1.41	1.22	1.04	1.02	0.68	0.75
	40	1.63	1.35	1.46	1.27	0.91	0.95	0.49	0.55
	100	1.74	1.42	1.48	1.31	0.77	0.85	0.34	0.39
40	20	1.86	1.35	1.64	1.28	1.02	0.96	0.55	0.58
	40	2.27	1.49	1.79	1.34	0.74	0.78	0.28	0.29
	100	2.78	1.66	1.79	1.39	0.44	0.55	0.10	0.10
100	20	2.41	1.42	2.03	1.32	1.02	0.89	0.46	0.44
	40	3.93	1.66	2.49	1.39	0.62	0.56	0.15	0.11
	100	7.76	2.02	2.55	1.39	0.19	0.20	0.01	0.01

Table 1
Continued.
c) $\phi = 0.50$

	$\Delta\delta$	0.00		0.20		0.50		0.80	
n_{orig}	n_{repl}	BF_{12}	BF_{13}	BF_{12}	BF_{13}	BF_{12}	BF_{13}	BF_{12}	BF_{13}
	20	2.31	1.50	2.02	1.43	1.17	1.09	0.59	0.69
20	40	2.72	1.63	2.19	1.51	0.96	1.00	0.37	0.47
	100	3.10	1.76	2.27	1.57	0.75	0.89	0.22	0.31
	20	3.51	1.63	2.79	1.51	1.21	1.02	0.46	0.51
40	40	5.33	1.87	3.38	1.63	0.82	0.80	0.19	0.22
	100	8.07	2.14	3.61	1.72	0.44	0.55	0.05	0.07
	20	6.02	1.76	4.32	1.59	1.38	0.93	0.40	0.36
100	40	16.30	2.14	7.04	1.73	0.84	0.57	0.11	0.08
	100	65.59	2.70	10.06	1.76	0.23	0.19	0.01	<0.01

d) $\phi = 0.80$

	$\Delta\delta$	0.00		0.20		0.50		0.80	
n_{orig}	n_{repl}	BF_{12}	BF_{13}	BF_{12}	BF_{13}	BF_{12}	BF_{13}	BF_{12}	BF_{13}
	20	8.45	2.05	6.13	1.91	2.05	1.34	0.61	0.69
20	40	12.88	2.29	7.75	2.07	1.67	1.22	0.33	0.46
	100	18.15	2.51	8.90	2.20	1.25	1.08	0.17	0.29
	20	24.77	2.29	14.64	2.08	3.03	1.25	0.59	0.50
40	40	72.13	2.72	26.25	2.31	2.29	0.98	0.21	0.20
	100	>100.00	3.19	40.58	2.48	1.26	0.66	0.04	0.06
	20	98.63	2.52	45.16	2.22	6.01	1.14	0.81	0.35
100	40	>100.00	3.19	>100.00	2.49	6.77	0.69	0.23	0.07
	100	>100.00	4.13	>100.00	2.59	3.19	0.23	0.01	<0.01

size determination methods (Maxwell et al., 2015) and limited resources. The remainder of this paper focuses only on studies in which both the original and replication study included 40 or 100 participants per group. Table 1 shows that studies with these samples allow for positive evidence in favor and against successful replication.

The second conclusion concerns the difference in performance of BF_{12} and BF_{13}. From Table 1 it can be inferred that BF_{12} is better suited to detect successful replication than BF_{13}. This is reflected by the higher values of BF_{12} than BF_{13} if $\Delta\delta = 0.00$ and 0.20. Furthermore, BF_{12} and BF_{13} do not show large differences in detecting non-successful replication. Thus, BF_{12} never performs worse than BF_{13} for values of ϕ, $\Delta\delta$ and n common in social and behavioral sciences that were presented in Table 1.

Figure 3 presents BF_{12} on a log-scale for $0.20 \leq \phi \leq 0.50$ and $n = 40, 100$. From Figure 3a it can be seen that $\frac{1}{3} \leq \mathrm{BF}_{12} \leq 3$ for $\phi = 0.20$ in most conditions. This implies that there is no positive evidence in favor of successful replication (H_1) nor for non-successful replication (H_2). The reason for this relatively weak evidence for one hypothesis, is that the two prior distributions $\pi(\delta_{repl}|H_1)$ and $\pi(\delta_{repl}|H_2)$ are very similar (see Figure 4). The replication data is not able to differentiate between the hypothesis of successful replication (H_1) and the hypothesis of non-successful replication (H_2) if $\phi = 0.20$.

The second conclusion that can be drawn from Figure 3 is that the strength of evidence in favor and against successful replication increases as ϕ changes from 0.20 through 0.35 to 0.5. For $\phi = 0.50$ (Figure 3c), it can be seen that if the sample sizes of the original and replication study are 100 participants per group, there is positive evidence for H_1 if $\Delta\delta < 0.3$ as $\mathrm{BF}_{12} < \frac{1}{3}$. On the other hand there is positive evidence for H_2 if $\Delta\delta > 0.5$. Only if $0.3 < \Delta\delta < 0.5$ the evidence is not sufficiently strong to decide convincingly in favor of either one of the two hypotheses. This area of unconvincing evidence is much larger if the sample size of the original or replication study decreases, and ranges from $0.2 < \Delta\delta < 0.7$ if $n_{orig} = n_{repl} = 40$.

The results presented in Table 1 and Figure 3 can be summarized as follows: if $\phi = 0.50$ is deemed a relevant difference in Cohen's d between the original and replication study, at least 100 participants per group should be included in both studies to acquire positive evidence for successful and non-successful replication. Note that both the choice of ϕ and the corresponding sample sizes that are required will be further discussed in the final section of this paper.

Data example: Mindset conditions and ambivalence

Data and sample size determination

The data example is taken from an original study performed by Henderson, de Liver, & Gollwitzer (2008) and a replication study by Lane & Gazarian (2015). This study investigated the ambivalence towards the statement that a list of sex offenders should be made publicly available for three experimental mindset conditions. The dependent variable "ambivalence" was measured on a scale from -3 to 3 where a higher score indicates a more ambivalent attitude. The participants were randomly assigned to one of three mindset conditions. In the first condition a one-sided mindset was activated, where participants were primed to rely on their own experiences. These experiences could be either in favor of or against the statement, but were intended to result in low ambivalence. In the second con-

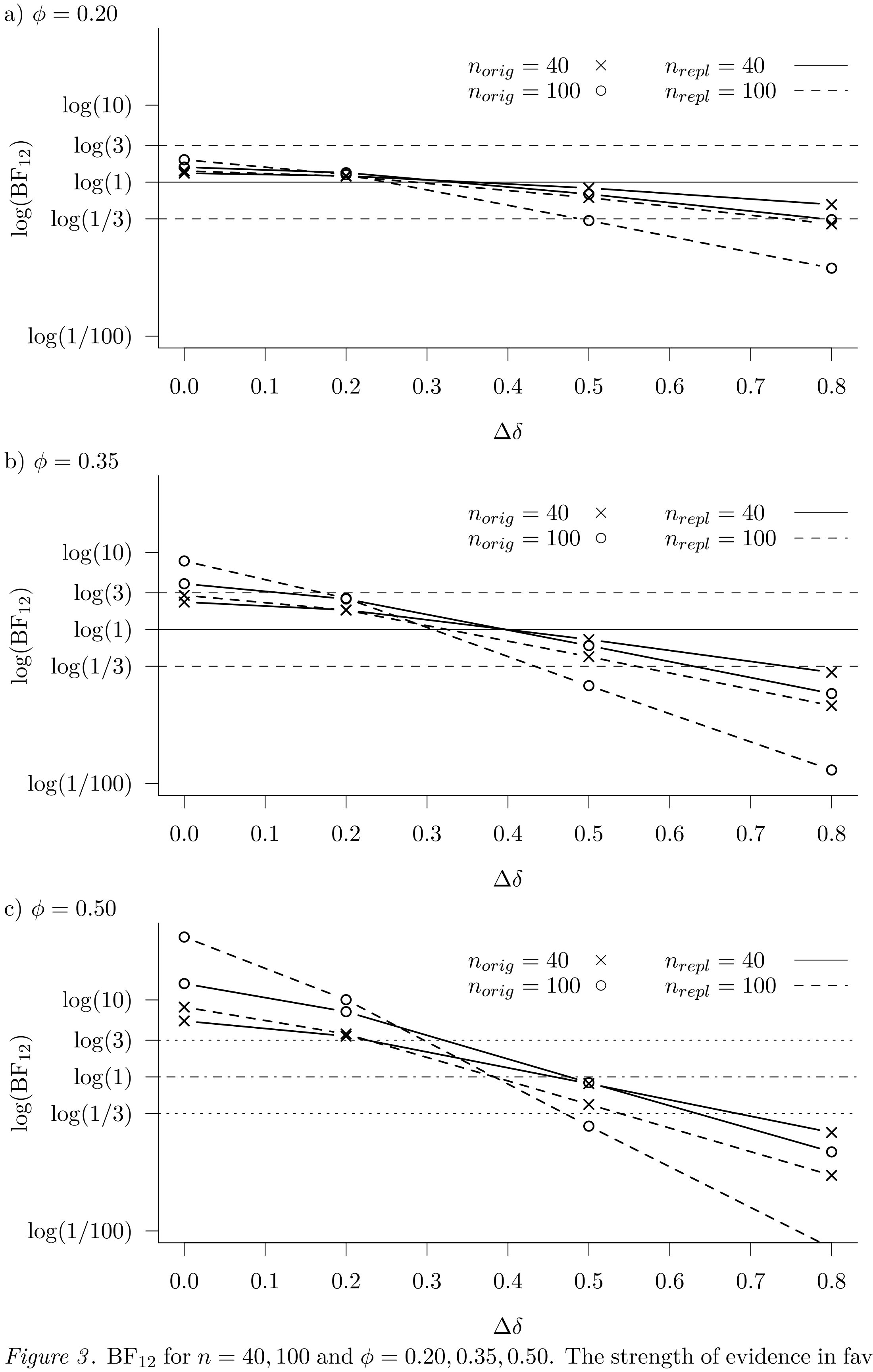
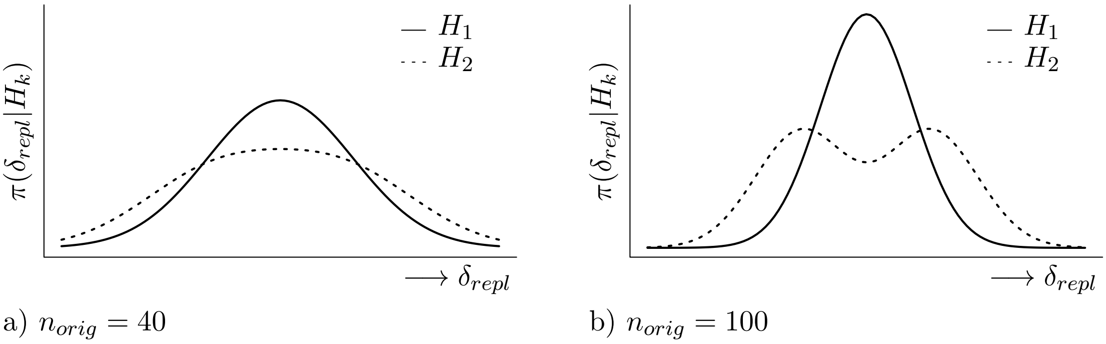


Figure 3. BF$_{12}$ for $n = 40, 100$ and $\phi = 0.20, 0.35, 0.50$. The strength of evidence in favor of successful replication (H_1) and non-successful replication under H_2 increases with sample sizes of the original and replication study and ϕ.



a) $n_{orig} = 40$

b) $n_{orig} = 100$

Figure 4. Very similar prior distributions $\pi(\delta_{repl}|H_k)$ for $k = 1, 2$ and $n_{orig} = 40, 100$ if $\phi = 0.20$.

dition a two-sided mindset was activated by stressing the importance of both perspectives. In the third, neutral, condition the mindset of participants was not purposefully activated. The ambivalence difference between the one- and two-sided mindset groups was of interest to the researchers, as well as the effect of the manipulations compared to the neutral condition. Therefore, not only an omnibus analysis of variance is conducted, but also all post-hoc t-tests were performed. In this situation the Bayes factors that were introduced in this paper can quantify the replication evidence for Cohen's d for each post-hoc t-test.

The replication study was an exact replication of the original study, and intended to achieve a minimum of 80% power in the omnibus analysis of variance test with $N \approx 90$, that is, 30 participants per group. With this sample size there is sufficient power to detect if Cohen's d differs from zero, but may not be sufficient to acquire positive evidence for replication of Cohen's d. If one wants to obtain evidence in favor of or against successful replication using the methods presented in this paper, the sample size should be determined differently. From the previous section it follow directly that if a difference of 0.50 in Cohen's d is deemed relevant ($\phi = 0.50$), the original study and the replication study should include 100 participants per group. If one would calculate BF_{12} for this original and replication data set, no positive evidence is expected to be found due to the very small sample sizes. In other words, due to the small sample sizes, values between $\frac{1}{3}$ and 3 are expected for BF_{12} and no conclusions can be drawn concerning successful or non-successful replication. The reanalysis of this data example will present Bayes factors for $\phi = 0.20, 0.35, 0.50, 0.80$ to show the properties of BF_{12} in this data example.

Open access data were retrieved for the replication study from `https://osf.io/79dey/`. It must be noted that the achieved sample size was smaller than the intended sample size, because participants were excluded from the analyses if they regularly interacted with a sexual offender. The data for the original study was generated based on the published sample summary statistics using BIEMS (Mulder et al., 2012). This resulted in data with exactly the same summary statistics as those that were used in the published article.

Table 2 shows the summary statistics and results of all post-hoc analyses. The p-values reported in Table 2 are based on the independent samples t-test with equal variance. The original study reported a significant difference in ambivalence score between the one-

Table 2

Summary Statistics and t-test Results for the Original and Replicated Data.

	Original study			Replication study		
Mindset condition	n	M	SD	n	M	SD
Neutral	16	1.23	1.64	24	-0.38	1.44
One-sided	15	0.16	1.85	23	-0.14	1.66
Two-sided	15	1.82	1.86	23	0.39	1.24
		δ	p		δ	p
One-sided vs. Two-sided		-0.93	.03		-0.38	.22
Neutral vs. One-sided		0.62	.11		-0.39	.61
Neutral vs. Two-sided		-0.34	.37		-0.59	.06

and two-sided mindset conditions ($p = .03$). The other two t-tests did not yield significant differences in ambivalence in the original study. Lane & Gazarian (2015) concluded that the results from the original study were not replicated, because the p-values for the main comparison (one-sided versus two-sided mindset) did not lead to similar conclusions. This conclusion is not justified for these data sets, as testing if Cohen's d in the original and replication study are different from zero, is not the same as testing if the two effect sizes are equal (Nieuwenhuis, Forstmann, & Wagenmakers, 2011). Furthermore, p-values do not indicate the strength of evidence (Wagenmakers, 2007). To properly quantify the evidence for (non-)replication the Bayes factors introduced in this paper were computed for this specific data set.

Results

For each post-hoc test BF_{12} was calculated using the steps in Figure 2. As Table 3 shows, all Bayes factors ≈ 1 for the group mean comparison of the one- and two-sided mindset conditions. For example BF_{12} for $\phi = 0.50$ indicates that the evidence for successful replication (H_1) is $\frac{1}{0.96} = 1.04$ times weaker than evidence for non-successful replication under H_2. These results indicate that there is no decisive evidence in favor of or against successful effect size replication for Cohen's d for the comparison between the one- and two-sided mindset conditions. In other words, irrespective of whether a difference in Cohen's d of 0.20, 0.35, 0.50, or 0.80 is deemed relevant, there is approximately equal evidence that δ_{repl} is similar to δ_{orig} or relevantly different from δ_{orig}. This conclusion was to be expected as the sample size was much smaller than the required sample size for our approach. Based on our approach it would not be possible to decide if the original study was successfully replicated, which deviates from the opinion of Lane & Gazarian (2015) who concluded that the original study was not successfully replicated.

The group mean comparison of the neutral and one-sided mindset conditions are all smaller than 1. This indicates that δ_{repl} was not consistent with δ_{orig}. However, for $0.20 \leq \phi \leq 0.50$, the Bayes factors do not provide positive evidence in terms of Kass & Raftery (1995). This may seem strange as the actual difference in Cohen's d between the original and replication study was $\Delta\delta = 0.62 + 0.39 = 1.01$ (see Table 2). This difference is rather large, so strong evidence against successful replication is expected to be found. Nonetheless, due to the small sample size the three prior distributions for δ_{repl} are very similar as was discussed in the previous section and Figure 4.

For the difference in ambivalence scores between the neutral and two-sided mindset conditions, the actual difference in Cohen's d estimates was -0.25. It can be seen that if a difference of $\phi = 0.20$ is deemed relevant, there Bayes factors are approximately 1. This indicates that the strength of evidence in favor of successful replication is similar to the strength of evidence for non-successful replication. However, if $\phi = 0.80$, BF_{12} increases to 4.29. This result indicates that there is positive evidence that Cohen's d was successfully replicated for the group mean difference between the neutral and two-sided conditions if differences in standardized mean difference are deemed relevant if they are 0.80. It must be noted that $\phi = 0.80$ seems too large in social and behavioral sciences research. For reasonable ranges of ϕ the sample sizes in this data example are too small to acquire positive evidence in favor of one of the competing hypotheses.

Table 3
Calculated BF_{12} Concerning the Data Example for $\phi = 0.20, 0.35, 0.50, 0.80$.

ϕ	one-sided vs. two-sided	neutral vs. one-sided	neutral vs. two-sided
0.20	0.97	0.77	1.09
0.35	1.01	0.75	1.32
0.50	0.96	0.41	1.77
0.80	1.29	0.31	4.29

Note: The Metropolis-Hastings-within-Gibbs sampler used to obtain samples from the posterior distributions in both the original and replication data set used 60,000 draws, including 10,000 draws for burn-in. The analysis procedures are in accordance with Appendices A and B.

Discussion

This paper introduced two Bayes factors that quantify the evidence for replication of Cohen's d. These Bayes factors, BF_{12} and BF_{13}, can be applied to any situation where the pairwise comparison of group means is of interest. BF_{12} is a more appropriate method to quantify the evidence for replication of Cohen's d than BF_{13}. BF_{12} weighs the relative evidence that δ_{repl} is in line with δ_{orig} (H_1) to the evidence that δ_{repl} deviates from δ_{orig} by a relevant shift ϕ in Cohen's d (H_2). The main conclusion of this paper is that if $\phi \geq 0.50$ is deemed relevant and approximately 100 participants per group are included in both the original and replication study, positive evidence can be found in favor of successful replication (H_1) for $0.3 < \Delta\delta$ and in favor of non-successful replication under H_2 for $\Delta\delta >$ 0.5.

However, the minimally relevant difference in Cohen's d between the original and replication study may be smaller than 0.50 in many situations. For example if $\delta_{orig} = 0.30$ and $\delta_{repl} = 0.10$ or 0.50, these effect sizes may already be interpreted as non-replication. Therefore, $\phi = 0.20$ may often be more appropriate in the context of social and behavioral sciences. As Table 1a and Figure 3a show for $\phi = 0.20$, even if $n = 100$ the evidence in favor of successful replication is very weak, i.e. $\text{BF}_{12} = 1.95$, and the area of unconvincing evidence for $\Delta\delta$ is large, i.e. $0.0 \leq \Delta\delta < 0.5$. The only way to decrease the area of unconvincing evidence is by increasing the sample size in the original and replication study. Additional analyses indicate that if both the original and replication study include 600 participants per group, $\mathrm{BF}_{12} = 54.88$ if $\Delta\delta = 0.00$ and $\mathrm{BF}_{12} = 0.27$ if $\Delta\delta = 0.20$. Thus, if a researcher deems differences of 0.20 in Cohen's d between the original and replication study relevant, then positive evidence in favor of successful and non-successful replication can be acquired if both studies include approximately 600 participants per group.

The sample sizes that are required to decide convincingly if Cohen's d in the original study has been replicated successfully are much higher than sample sizes that are commonly used in the social and behavioral sciences. This result is in line with Maxwell, Lau, & Howard (2015) who state that researchers often use flawed methods in sample size determination that underestimated the required sample size for replication studies. For example the method of Maxwell et al. (2015) as was introduced in the introduction requires large samples too. Their approach uses confidence intervals for $\Delta\delta$ and tests equivalence ($\Delta\delta \approx 0$) and nonequivalence ($\Delta\delta \not\approx 0$) of the difference in Cohen's d in two tests. The test for equivalence requires more than 1,000 participants per group, due to the large sampling variation in $\Delta\delta$ (Maxwell, 2013; Maxwell et al., 2015). To test for nonequivalence of Cohen's d in the original and replication study, Maxwell (2013) suggests that possibly 200 participants per group in the original study and 65 participants per group in the replication study could be sufficient. The results of this paper and Maxwell et al. (2015) highlighted that large samples are needed to clarify what scientific findings should become part of theory and what results should remain only anecdotal.

It is clear that deciding whether Cohen's d was successfully replicated is not straightforward. Researchers who set up a replication study should carefully determine the required sample size based on the original study sample size and their choice of ϕ. We advice researchers to carefully consider what value of ϕ is appropriate in their specific situation, based on the norms in their field of study and the impact of the study results. The required sample size can be determined using the software that accompanies this paper. First one should create data sets with diverse n_{repl} in BIEMS (Mulder et al., 2012) and calculate BF_{12} using the `R` function or `Shiny` package. These Bayes factors can be plotted similarly to Figure 3 to decide how many participants should be included to acquire convincing evidence in favor of successful and non-successful replication.

References

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

Bayarri, M. J., & Mayoral, A. M. (2002). Bayesian design of "successful" replications. *The American Statistician*, *56*(3), 207–214. doi:10.1198/000313002155.

Behboodian, J. (1970). On a mixture of normal distributions. *Biometrika*, *57*(1), 215–217. doi:10.1093/biomet/57.1.215.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2016). *shiny: Web Application Framework for R.* R package version 0.13.1. URL: `http://CRAN.R-project.org/package=shiny`

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*(432), 1313–1321. doi:10.1080/01621459.1995.10476635.

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the MetropolisâĂŞHastings output. *Journal of the American Statistical Association*, *96*(453), 270–281. doi:10.1198/016214501750332848.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi:10.1037/0033-2909.112.1.155.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286–300. doi:10.1111/j.1745-6924.2008.00079.x.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*(4), 532–574. doi:10.1177/0013164401614002.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: psychology. *PLoS ONE*, *11*(2), 1–12. doi:10.1371/journal.pone.0149794.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis (2nd ed.).* Boca Raton, FL: Chapman and Hall/CRC.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2016). *mvtnorm: Multivariate Normal and t Distributions.* R package version 1.0-5. URL: `http://CRAN.R-project.org/package=mvtnorm`

Henderson, M., de Liver, Y., & Gollwitzer, P. (2008). The effects of an implemental mindset on attitude strength. *Journal of Personality and Social Psychology*, *94*(3), 396–411. doi:10.1037/0022-3514.94.3.396.

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

Lane, K. A., & Gazarian, D. (2015). Replication of Henderson, de Liver, & Gollwitzer (2008, Journal of Personality and Social Psychology, experiment 5). URL: `osf.io/79dey`

Maxwell, S. (2013). *Methodologial issues in planning and interpreting replication studies.* Powerpoint presentation at American Psychological Association. Honolulu, Hawaii.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. doi:10.1037/a0039400.

Mulaik, S. A., Raju, N. S., & Harhman, R. A. (2016). There is a time and place for significance testing. In L. S. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.) *What If There Were No Significance Tests?*, (pp. 61–106). New York, NY: Routledge, classic ed.

Mulder, J., Hoijtink, H., & Leeuw, C. (2012). BIEMS: A Fortran 90 program for calculating bayes factors for inequality and equality constrained models. *Journal of Statistical Software*, *46*(1), 1–39. doi:10.18637/jss.v046.i02.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107. doi:10.1038/nn.2886.

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

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: `http://www.R-project.org/`

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80. doi:10.1177/1745691613514755.

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

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. doi:10.1037/a0036731.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p*values. *Psychonomic bulletin & review*, *14*(5), 779–804. doi:10.3758/bf03194105.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. doi:10.1177/1745691612463078.

Appendix A
Obtaining the posterior distribution of δ_{orig}

To obtain the posterior distribution of $\delta_{orig} = N(\bar{\delta}_{orig}, \gamma^2)$ that summarizes the information on Cohen's d and its uncertainty in the original data set, non-informative priors are used for μ_1, δ_{orig}, and σ^2: $\pi(\mu_1) = \pi(\delta_{orig}) = N(0, 10000), \pi(\sigma^2) = \Gamma^{-1}(0.001, 0.001)$. Furthermore, the density of the data as provided in Equation 4 is needed.

A Gibbs sampler with Metropolis-Hastings-step (for more information, see Gelman et al., 2004, Chapter 11) for the σ^2 is implemented with the goal to obtain the mean and variance of the posterior distribution of δ_{orig}, that is $\bar{\delta}_{orig}$ and γ^2. Values are sampled iteratively from the conditional posterior distributions $p(\mu_1|\boldsymbol{y}, \boldsymbol{D}, \delta_{orig}, \sigma^2)$, $p(\delta_{orig}|\boldsymbol{y}, \boldsymbol{D}, \mu_1, \sigma^2)$, and $p(\sigma^2|\boldsymbol{y}, \boldsymbol{D}, \mu_1, \delta_{orig})$:

$$p(\mu_1|\boldsymbol{y}, \boldsymbol{D}, \delta_{orig}, \sigma^2) = N(\frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i + D_i\delta\sigma)(\frac{N}{\sigma^2} + \frac{1}{10000})^{-1}, (\frac{N}{\sigma^2} + \frac{1}{10000})^{-1}) \ , \tag{A1}$$

$$p(\delta_{orig}|\boldsymbol{y}, \boldsymbol{D}, \mu_1, \sigma^2) = N(\frac{1}{\sigma}\sum_{i=1}^{N} D_i(y_i - \mu_1)(n_2 + \frac{1}{10000})^{-1}, (n_2 + \frac{1}{10000})^{-1}) \ , \tag{A2}$$

and

$$p(\sigma^2|\boldsymbol{y}, \boldsymbol{D}, \mu_1, \delta_{orig}) \propto (\sigma^2)^{-\frac{N}{2}-0.001-1} e^{-\frac{1}{\sigma^2}(0.001+\frac{1}{2}\sum_{i=1}^{N}(y_i-\mu_1+D_i\cdot\delta_{orig}\cdot\sigma)^2)} \ . \tag{A3}$$

The conditional posterior distribution of σ^2 is of unknown form. To be able to sample from this conditional posterior distribution a Metropolis-Hastings-step was implemented, using a non-adaptive proposal distribution:

$$\begin{aligned} q(\sigma^2|\boldsymbol{y}, \boldsymbol{D}) &\propto (\sigma^2)^{-\frac{N}{2}-0.001-1} e^{-\frac{1}{\sigma^2}(0.001+\frac{1}{2}\sum_{i=1}^{N}(y_i-m_1+D_i\cdot d\cdot s)^2)} \\ &= \Gamma^{-1}(0.001 + \frac{N}{2}, 0.001 + \frac{1}{2}\sum_{i=1}^{N}(y_i - m_1 + D_i \cdot d \cdot s)^2) \ , \end{aligned} \tag{A4}$$

where m_1 represents the mean score of the first group in the sample, d is the estimate of Cohen's d, and s is the estimated pooled standard deviation in the original study. The iterative process results in a sample from the posterior distribution of δ_{orig} containing the values $\delta_{orig(t)}$ for iteration $t = 1, 2, \ldots, T$. These samples result in estimates of $\bar{\delta}_{orig} = T^{-1}\sum_{t=1}^{T}\delta_{orig(t)}$ and $\gamma^2 = T^{-1}\sum_{t=1}^{T}(\delta_{orig(t)} - \bar{\delta}_{orig})^2$.

Appendix B
Obtaining the posterior distributions of μ_1, δ_{repl}, and σ^2

To sample μ_1, δ_{repl}, and σ^2 from their respective conditional posterior distributions, four densities are needed: the density of the data as defined in Equation 4, the prior distribution for δ_{repl} under the hypothesis of interest as displayed in Equations 1-3, and the non-informative prior distributions on μ_1 and σ^2 as used in Appendix A. The sampling procedure is similar to the one used in the analysis of the original data (see Appendix A). The conditional posterior distributions for μ_1 and σ^2 are identical to the conditional posterior distributions that were presented in Equations A1 and A3 in Appendix A, but now based on the replication data set. The Metropolis-Hastings-step in analyzing the replication data uses the same non-adaptive proposal distribution as displayed in Equation A4, where the sample mean, Cohen's d, variance, and data now refer to the replication data set.

The conditional posterior distribution of δ_{repl} is a normal distribution under H_1 and H_3. For the first hypothesis δ_{repl} is sampled from the following normal distribution in each iteration:

$$p(\delta_{repl}|\boldsymbol{y}, \boldsymbol{D}, \mu_1, \sigma^2, H_1) \propto N((\frac{\bar{\delta}_{orig}}{\gamma^2} - \frac{1}{\sigma}\sum_{i=1}^{N} D_i(y_i - \mu_1))(n_2 + \frac{1}{\gamma^2})^{-1}, (n_2 + \frac{1}{\gamma^2})^{-1}) \ . \quad \text{(B1)}$$

The conditional posterior distribution under the third hypothesis is:

$$p(\delta_{repl}|\boldsymbol{y}, \boldsymbol{D}, \mu_1, \sigma^2, H_3) \propto N((\frac{\bar{\delta}_{orig}}{\gamma^2 + \phi^2} - \frac{1}{\sigma}\sum_{i=1}^{N} D_i(y_i - \mu_1))(n_2 + \frac{1}{\gamma^2 + \phi^2})^{-1}, (n_2 + \frac{1}{\gamma^2 + \phi^2})^{-1}) \ . \quad \text{(B2)}$$

The second hypothesis is evaluated using a separate sampling process for each element of the prior distribution on δ_{repl}:

$$p(\delta_{repl}|\boldsymbol{y}, \boldsymbol{D}, \mu_1, \sigma^2, \pi_{min}) \propto N((\frac{\bar{\delta}_{orig} - \phi}{\gamma^2} - \frac{1}{\sigma}\sum_{i=1}^{N} D_i(y_i - \mu_1))(n_2 + \frac{1}{\gamma^2})^{-1}, (n_2 + \frac{1}{\gamma^2})^{-1}) \ , \quad \text{(B3)}$$

and

$$p(\delta_{repl}|\boldsymbol{y}, \boldsymbol{D}, \mu_1, \sigma^2, \pi_{max}) \propto N((\frac{\bar{\delta}_{orig} + \phi}{\gamma^2} - \frac{1}{\sigma}\sum_{i=1}^{N} D_i(y_i - \mu_1))(n_2 + \frac{1}{\gamma^2})^{-1}, (n_2 + \frac{1}{\gamma^2})^{-1}) \ . \quad \text{(B4)}$$

For each hypothesis this iterative process renders three samples of values for μ_1, δ_{repl}, and σ^2 from their posterior distributions: $\mu_{1(g)}, \delta_{repl(g)}, \sigma^2_{(g)}$ for iteration $g = 1, 2, \ldots, G$. The posterior means of these distributions are selected to be the assessment values μ_{1*}, δ_{repl*}, and σ^2_* in Equation 6. Note that $\mu_{1*} = G^{-1}\sum_{g=1}^{G} \mu_{1(g)}$, $\bar{\delta}_{repl*} = G^{-1}\sum_{g=1}^{G} \delta_{repl(g)}$, and $\sigma^2_* = G^{-1}\sum_{g=1}^{G} \sigma^2_{(g)}$. Because $\pi(\delta_{repl}|H_k)$ differs per hypothesis, the assessment values are different for each hypotheses.

Appendix C
Computing the marginal likelihood of the replication data given H_1, H_2, or H_3

As can be seen from Equation 5 the marginal likelihood of the data under the competing hypotheses is required to compute BF_{12} and BF_{13}. Chib (1995) and Chib & Jeliazkov (2001) proposed to estimate the natural logarithm of the marginal of the data using Equation 6. This Appendix discusses the computation of the fifth and sixth component in Equation 6.

Joint conditional density for the first group mean and Cohen's d: $p(\mu_{1*}, \delta_{repl*}|\boldsymbol{y}, \boldsymbol{D}, \sigma^2_*)$. The joint conditional posterior distribution of μ_1 and δ_{repl} is a bivariate normal distribution:

$$p(\mu_{1*}, \delta_{repl*}|\boldsymbol{y}, \boldsymbol{D}, \sigma^2_*) = N\left(\begin{bmatrix} M_{\mu_1} \\ M_{\delta_{repl}} \end{bmatrix}, \begin{bmatrix} \frac{N}{\sigma^2_*} + \frac{1}{10000} & -\frac{n_2}{\sigma_*} \\ -\frac{n_2}{\sigma_*} & n_2 + \frac{1}{\tau_0^2} \end{bmatrix}^{-1}\right) \quad \text{for } H_1, H_3 \ , \tag{C1}$$

where τ_0^2 is the prior variance of δ_{repl} under each hypothesis: γ^2 for H_1 and $\gamma^2 + \phi$ for H_3. Under the second hypothesis, this equation decomposes into

$$p(\mu_{1*}, \delta_{repl*}|\boldsymbol{y}, \boldsymbol{D}, \sigma^2_*) = \frac{1}{2}p(\mu_{1*}, \delta_{repl*}|\boldsymbol{y}, \boldsymbol{D}, \sigma^2_*, \pi_{min}) + \frac{1}{2}p(\mu_{1*}, \delta_{repl*}|\boldsymbol{y}, \boldsymbol{D}, \sigma^2_*, \pi_{max}) \ ,$$

where for each element $\tau_0^2 = \gamma^2$. The posterior means M_{μ_1} and $M_{\delta_{repl}}$ are

$$\begin{aligned}
M_{\mu_1} &= (1 - Bn_2)^{-1}[A - B \cdot n_2^2 \cdot m_2) \\
M_{\delta_{repl}} &= (\frac{\delta_0}{\tau_0^2} - \frac{1}{\sigma_*}\sum_{i=1}^{N} D_i(y_i - M_{\mu_1}))(n_2 + \frac{1}{\tau_0^2})^{-1} \ \text{, where} \\
A &= [\frac{1}{\sigma^2_*}\sum_{i=1}^{N} y_i + \frac{1}{\sigma^2_*}\sum_{i=1}^{N}(D_i\sigma_*\frac{\delta_0}{\tau_0^2}(n_2 + \frac{1}{\tau_0^2})^{-1})][\frac{N}{\sigma^2_*} + \frac{1}{10000}]^{-1} \ \text{and} \\
B &= \frac{1}{\sigma^2_*}(\frac{N}{\sigma^2_*} + \frac{1}{10000})^{-1}(n_2 + \frac{1}{\tau_0^2})^{-1}
\end{aligned}$$

for the hypotheses used in this paper. The prior mean of δ_{repl} is denoted by δ_0. For H_1 and H_3, $\delta_0 = \bar{\delta}_{orig}$. For both components of H_2, δ_0 equals $\bar{\delta}_{orig} - \phi$ and $\bar{\delta}_{orig} + \phi$ respectively. Note that m_2 is the sample mean of Group 2 in the replication data set. The density of this multivariate normal distribution was evaluated using the `R` package `mvtnorm` (Genz et al., 2016).

Marginal density of the variance component: $p(\sigma^2_*|\boldsymbol{y}, \boldsymbol{D})$. The density of $p(\sigma^2_*|\boldsymbol{y}, \boldsymbol{D})$ is of unknown form. Applied to our situation, Chib & Jeliazkov (2001) proposed to estimate this density as follows:

$$p(\sigma^2_*|\boldsymbol{y}, \boldsymbol{D}, H_k) = \frac{G^{-1}\sum_{g=1}^{G}\alpha(\sigma^2_{(g)}, \sigma^2_*) \cdot q(\sigma^2_*|\boldsymbol{y}, \boldsymbol{D}, \mu_{1(g)}, \delta_{(g)})}{G^{-1}\sum_{g=1}^{G}\alpha(\sigma^2_*, \sigma^2_{c(g)})} \ , \tag{C2}$$

where $\alpha(\sigma^2_{(a)}, \sigma^2_{(b)}) = \min(1, \frac{p(\sigma^2_{(b)}|\boldsymbol{y},\boldsymbol{D},\mu_{1(a)},\delta_{repl(a)},H_k)}{p(\sigma^2_{(a)}|\boldsymbol{y},\boldsymbol{D},\mu_{1(a)},\delta_{repl(a)},H_k)}\frac{q(\sigma^2_{(a)}|\boldsymbol{y},\boldsymbol{D},\mu_{1(a)},\delta_{repl(a)},H_k)}{q(\sigma^2_{(b)}|\boldsymbol{y},\boldsymbol{D},\mu_{1(a)},\delta_{repl(a)},H_k)})$ denotes the probability of selecting $\sigma^2_{(b)}$ over $\sigma^2_{(a)}$ as a draw from the posterior distribution. Note that $\sigma^2_{(g)}$ refers to the sampled value of σ^2 in iteration g (see Equation A3), while $\sigma^2_{c(g)}$ refers to the candidate for σ^2 that was sampled from the proposal distribution in iteration g (see Equation A4).