Replication of "null results" – Absence of evidence or evidence of absence?

Samuel Pawel[1*†], **Rachel Heyard**[1†], **Charlotte Micheloud**[1], **Leonhard Held**[1]

***For correspondence:** samuel.pawel@uzh.ch (SP)

[†]Contributed equally

[1]Epidemiology, Biostatistics and Prevention Institute, Center for Reproducible Science, University of Zurich, Switzerland

Abstract In several large-scale replication projects, statistically non-significant results in both the original and the replication study have been interpreted as a "replication success". Here we discuss the logical problems with this approach: Non-significance in both studies does not ensure that the studies provide evidence for the absence of an effect and "replication success" can virtually always be achieved if the sample sizes are small enough. In addition, the relevant error rates are not controlled. We show how methods, such as equivalence testing and Bayes factors, can be used to adequately quantify the evidence for the absence of an effect and how they can be applied in the replication setting. Using data from the Reproducibility Project: Cancer Biology we illustrate that many original and replication studies with "null results" are in fact inconclusive, and that their replicability is lower than suggested by the non-significance approach. We conclude that it is important to also replicate studies with statistically non-significant results, but that they should be designed, analyzed, and interpreted appropriately.

Introduction

Absence of evidence is not evidence of absence – the title of the 1995 paper by Douglas Altman and Martin Bland has since become a mantra in the statistical and medical literature (***Altman and Bland, 1995***). Yet, the misconception that a statistically non-significant result indicates evidence for the absence of an effect is unfortunately still widespread (***Makin and de Xivry, 2019***). Such a "null result" – typically characterized by a *p*-value of $p > 0.05$ for the null hypothesis of an absent effect – may also occur if an effect is actually present. For example, if the sample size of a study is chosen to detect an assumed effect with a power of 80%, null results will incorrectly occur 20% of the time when the assumed effect is actually present. If the power of the study is lower, null results will occur more often. In general, the lower the power of a study, the greater the ambiguity of a null result. To put a null result in context, it is therefore critical to know whether the study was adequately powered and under what assumed effect the power was calculated (***Hoenig and Heisey, 2001***; ***Greenland, 2012***). However, if the goal of a study is to explicitly quantify the evidence for the absence of an effect, more appropriate methods designed for this task, such as equivalence testing (***Senn, 2008***; ***Wellek, 2010***; ***Lakens, 2017***) or Bayes factors (***Kass and Raftery, 1995***; ***Goodman, 1999***), should be used from the outset.

The interpretation of null results becomes even more complicated in the setting of replication studies. In a replication study, researchers attempt to repeat an original study as closely as possible in order to assess whether consistent results can be obtained with new data (***National Academies of Sciences, Engineering, and Medicine, 2019***). In the last decade, various large-scale replication

projects have been conducted in diverse fields, from the biomedical to the social sciences (***Prinz et al., 2011***; ***Begley and Ellis, 2012***; ***Klein et al., 2014***; ***Open Science Collaboration, 2015***; ***Camerer et al., 2016***, ***2018***; ***Klein et al., 2018***; ***Cova et al., 2018***; ***Errington et al., 2021***, among others). Most of these projects reported alarmingly low replicability rates across a broad spectrum of criteria for quantifying replicability. While most of these projects restricted their focus on original studies with statistically significant results ("positive results"), the *Reproducibility Project: Psychology* (RPP, ***Open Science Collaboration, 2015***), the *Reproducibility Project: Experimental Philosophy* (RPEP, ***Cova et al., 2018***), and the *Reproducibility Project: Cancer Biology* (RPCB, ***Errington et al., 2021***) also attempted to replicate some original studies with null results – either non-significant or interpreted as showing no evidence for a meaningful effect by the original authors.

While the RPEP and RPP interpreted non-significant results in both original and replication study as a "replication success" for some individual replications (see, for example, the replication of ***McCann*** (***2005***, replication report: https://osf.io/wcm7n) or the replication of ***Ranganath and Nosek*** (***2008***, replication report: https://osf.io/9xt25)), they excluded the original null results in the calculation of an overall replicability rate based on significance. In contrast, the RPCB explicitly defined null results in both the original and the replication study as a criterion for "replication success". According to this criterion, $11/15 = 73\%$ replications of original null effects were successful. Four additional criteria were used to assess successful replications of original null results: comparing original effect size with a 95% confidence interval of the replication effect size, comparing replication effect size with a 95% confidence interval of the original effect size, comparing replication effect size with a 95% prediction interval of the original effect size, and combining the original and replication effect sizes in meta-analysis. The suitability of these criteria in the context of replications of original null effects will be discussed in our conclusion.

RH: listed. but some the significance criterion comments are valid also for some of those criteria. Should we here say that we will discuss them in the conclusion, or mention it here already? - maybe delete the sentence "The suitability..."

We believe that there are several logical problems with the "non-significance" criterion. First, if the original study had low statistical power, a non-significant result is highly inconclusive and does not provide evidence for the absence of an effect. It is then unclear what exactly the goal of the replication should be – to replicate the inconclusiveness of the original result? On the other hand, if the original study was adequately powered, a non-significant result may indeed provide some evidence for the absence of an effect when analyzed with appropriate methods, so that the goal of the replication is clearer. However, the criterion by itself does not distinguish between these two cases. Second, with this criterion researchers can virtually always achieve replication success by conducting a replication study with a very small sample size, such that the *p*-value is non-significant and the result are inconclusive. This is because the null hypothesis under which the *p*-value is computed is misaligned with the goal of inference, which is to quantify the evidence for the absence of an effect. We will discuss methods that are better aligned with this inferential goal. Third, the criterion does not control the error of falsely claiming the absence of an effect at some predetermined rate. This is in contrast to the standard replication success criterion of requiring significance from both studies (also known as the two-trials rule, see Section 12.2.8 in ***Senn, 2008***), which ensures that the error of falsely claiming the presence of an effect is controlled at a rate equal to the squared significance level (for example, $5\% \times 5\% = 0.25\%$ for a 5% significance level). The non-significance criterion may be intended to complement the two-trials rule for null results. However, it fails to do so in this respect, which may be required by regulators and funders.

The aim of this paper is to present two principled approaches for analyzing replication studies of null results, which can address the limitations of the non-significance criterion. In the following, we will use the null results replicated in the RPCB to illustrate the problems of the non-significance criterion. We then explain and illustrate how both frequentist equivalence testing and Bayesian hypothesis testing can be used to overcome them. It is important to note that it is not our intent

to diminish the enormously important contributions of the RPCB, but rather to build on their work and provide recommendations for future replication researchers.

Null results from the Reproducibility Project: Cancer Biology

Figure 1 shows effect estimates on standardized mean difference (SMD) scale with 95% confidence intervals from two RPCB study pairs. In both study pairs, the original and replications studies are "null results" and therefore meet the non-significance criterion for replication success (the two-sided *p*-values are greater than 0.05 in both the original and the replication study). However, intuition would suggest that the conclusions in the two pairs are very different.

The original study from ***Dawson et al.*** (***2011***) and its replication both show large effect estimates in magnitude, but due to the very small sample sizes, the uncertainty of these estimates is large, too. With such low sample sizes, the results seem inconclusive. In contrast, the effect estimates from ***Goetz et al.*** (***2011***) and its replication are much smaller in magnitude and their uncertainty is also smaller because the studies used larger sample sizes. Intuitively, the results seem to provide more evidence for a zero (or negligibly small) effect. While these two examples show the qualitative difference between absence of evidence and evidence of absence, we will now discuss how the two can be quantitatively distinguished.

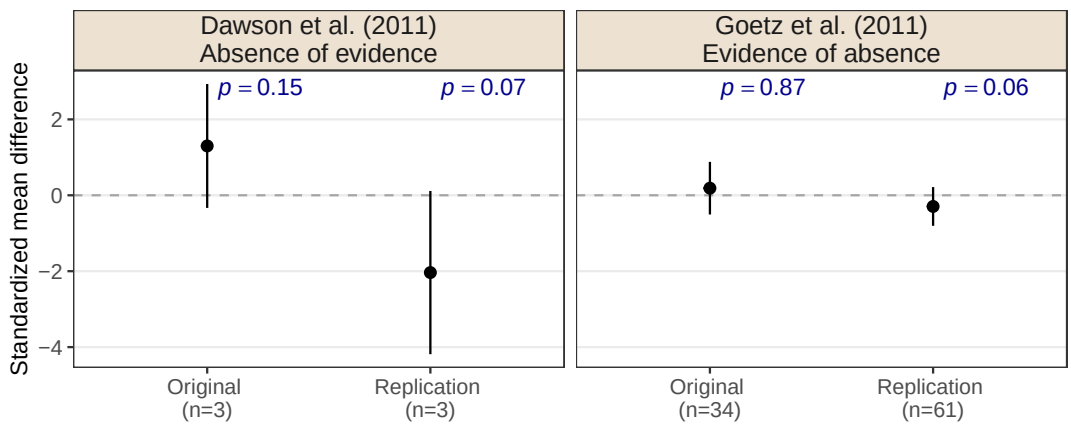

Figure 1. Two examples of original and replication study pairs which meet the non-significance replication success criterion from the Reproducibility Project: Cancer Biology (***Errington et al., 2021***). Shown are standardized mean difference effect estimates with 95% confidence intervals, sample sizes n, and two-sided *p*-values p for the null hypothesis that the effect is absent.

Methods for assessing replicability of null results

There are both frequentist and Bayesian methods that can be used for assessing evidence for the absence of an effect. ***Anderson and Maxwell*** (***2016***) provide an excellent summary in the context of replication studies in psychology. We now briefly discuss two possible approaches – frequentist equivalence testing and Bayesian hypothesis testing – and their application to the RPCB data.

Frequentist equivalence testing

Equivalence testing was developed in the context of clinical trials to assess whether a new treatment – typically cheaper or with fewer side effects than the established treatment – is practically equivalent to the established treatment (***Wellek, 2010***). The method can also be used to assess whether an effect is practically equivalent to an absent effect, usually zero. Using equivalence testing as a way to put non-significant results into context has been suggested by several authors

(***Hauck and Anderson, 1986***; ***Campbell and Gustafson, 2018***). The main challenge is to specify the margin $\Delta > 0$ that defines an equivalence range $[-\Delta, +\Delta]$ in which an effect is considered as absent for practical purposes. The goal is then to reject the null hypothesis that the true effect is outside the equivalence range. This is in contrast to the usual null hypothesis of superiority tests which state that the effect is zero or smaller than zero, see Figure 2 for an illustration.

To ensure that the null hypothesis is falsely rejected at most $\alpha \times 100\%$ of the time, the standard approach is to declare equivalence if the $(1 - 2\alpha) \times 100\%$ confidence interval for the effect is contained within the equivalence range, for example, a 90% confidence interval for $\alpha = 5\%$ (***Westlake, 1972***). This procedure is equivalent to declaring equivalence when two one-sided tests (TOST) for the null hypotheses of the effect being greater/smaller than $+\Delta$ and $-\Delta$, are both significant at level α (***Schuirmann, 1987***). A quantitative measure of evidence for the absence of an effect is then given by the maximum of the two one-sided *p*-values (the TOST *p*-value). A reasonable criterion for replication success of original null results may therefore be to require that both the original and the replication TOST *p*-values are smaller than some level α (conventionally 0.05). Equivalently, the criterion would require the $(1 - 2\alpha) \times 100\%$ confidence intervals of the original and the replication to be included in the equivalence region. In contrast to the non-significance criterion, this criterion controls the error of falsely claiming replication success at level α^2 when there is a true effect outside the equivalence margin, thus complementing the usual two-trials rule in drug regulation (***Senn, 2008***, section 12.2.8).

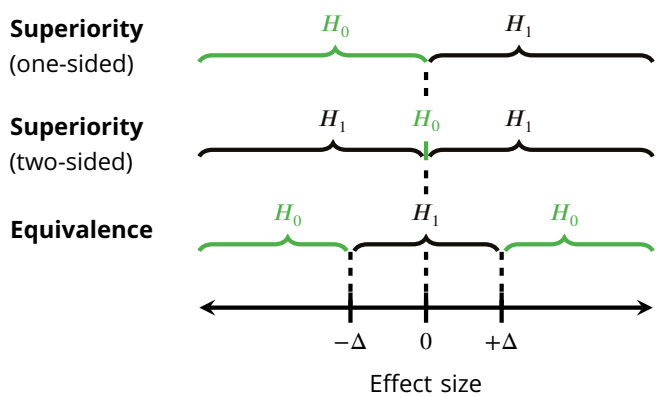

Figure 2. Null hypothesis (H_0) and alternative hypothesis (H_1) for superiority and equivalence tests (with equivalence margin $\Delta > 0$).

Returning to the RPCB data, Figure 3 shows the standardized mean difference effect estimates with 90% confidence intervals for all 15 effects which were treated as null results by the RPCB.[1] Most of them showed non-significant *p*-values ($p > 0.05$) in the original study. It is, however, noteworthy that two effects from the second experiment from the original paper 48 were regarded as null results despite their statistical significance. According to the non-significance criterion (requiring $p > 0.05$ in original and replication study), there are 11 "successes" out of total 15 null effects, as reported in Table 1 from ***Errington et al.*** (***2021***).

We will now apply equivalence testing to the RPCB data. The dotted red lines in Figure 3 represent an equivalence range for the margin $\Delta = 0.74$, which ***Wellek*** (***2010***, Table 1.1) classifies as "liberal". However, even with this generous margin, only 4 of the 15 study pairs are able to establish replication success at the 5% level, in the sense that both the original and the replication 90% confidence interval fall within the equivalence range (or, equivalently, that their TOST *p*-values are

[1] There are four original studies with null effects for which two or three "internal" replication studies were conducted, leading in total to 20 replications of null effects. As in the RPCB main analysis (***Errington et al., 2021***), we aggregated their SMD estimates into a single SMD estimate with fixed-effect meta-analysis and recomputed the replication *p*-value based on a normal approximation. For the original studies and the single replication studies we report the *p*-values as provided by the RPCB.

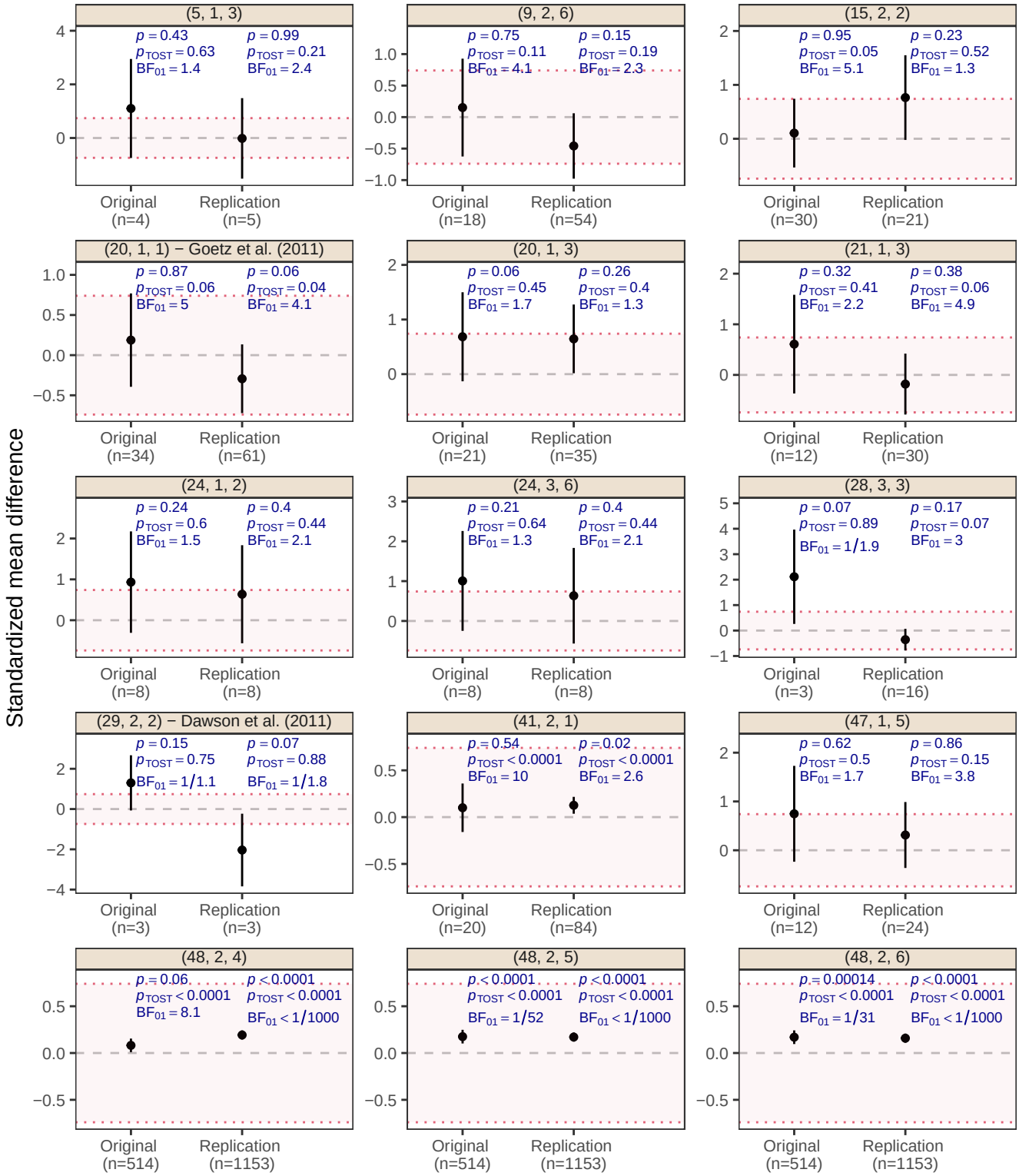

Figure 3. Effect estimates on standardized mean difference (SMD) scale with 90% confidence interval for the "null results" and their replication studies from the Reproducibility Project: Cancer Biology (***Errington et al., 2021***). The identifier above each plot indicates (original paper number, experiment number, effect number). Two original effect estimates from original paper 48 were statistically significant at $p < 0.05$, but were interpreted as null results by the original authors and therefore treated as null results by the RPCB. The two examples from Figure 1 are indicated in the plot titles. The dashed gray line represents the value of no effect (SMD $= 0$), while the dotted red lines represent the equivalence range with a margin of $\Delta = 0.74$, classified as "liberal" by ***Wellek*** (***2010***, Table 1.1). The *p*-value p_{TOST} is the maximum of the two one-sided *p*-values for the null hypotheses of the effect being greater/less than $+\Delta$ and $-\Delta$, respectively. The Bayes factor BF_{01} quantifies the evidence for the null hypothesis $H_0: \text{SMD} = 0$ against the alternative $H_1: \text{SMD} \neq 0$ with normal unit-information prior assigned to the SMD under H_1.

smaller than 0.05). For the remaining 11 studies, the situation remains inconclusive and there is no evidence for the absence or the presence of the effect. For instance, the previously discussed example from ***Goetz et al.*** (***2011***) marginally fails the criterion ($p_{\text{TOST}} = 0.06$ in the original study and $p_{\text{TOST}} = 0.04$ in the replication), while the example from ***Dawson et al.*** (***2011***) is a clearer failure ($p_{\text{TOST}} = 0.75$ in the original study and $p_{\text{TOST}} = 0.88$ in the replication) as both effect estimates even lie outside the equivalence margin.

The post-hoc specification of equivalence margins is controversial. Ideally, the margin should be specified on a case-by-case basis in a pre-registered protocol before the studies are conducted by researchers familiar with the subject matter. In the social and medical sciences, the conventions of ***Cohen*** (***1992***) are typically used to classify SMD effect sizes (SMD $= 0.2$ small, SMD $= 0.5$ medium, SMD $= 0.8$ large). While effect sizes are typically larger in preclinical research, it seems unrealistic to specify margins larger than 1 on SMD scale to represent effect sizes that are absent for practical purposes. It could also be argued that the chosen margin $\Delta = 0.74$ is too lax compared to margins commonly used in clinical research (***Senn, 2008***, chapter 22). However, as illustrated in Figure 4 from the sensitivity analysis in our appendix, for realistic margins between 0 and 1, the proportion of replication successes remains below 50% for the conventional $\alpha = 0.05$ level. To achieve a success rate of $11/15 = 73\%$, as was achieved with the non-significance criterion from the RPCB, unrealistic margins of $\Delta > 2$ are required.

Bayesian hypothesis testing

The distinction between absence of evidence and evidence of absence is naturally built into the Bayesian approach to hypothesis testing. A central measure of evidence is the Bayes factor (***Kass and Raftery, 1995***), which is the updating factor of the prior odds to the posterior odds of the null hypothesis H_0 versus the alternative hypothesis H_1

$$\text{Posterior odds} = \text{Prior odds} \times \text{Bayes factor } \mathrm{BF}_{01}\,.$$

The Bayes factor quantifies how much the observed data have increased or decreased the probability of the null hypothesis H_0 relative to the alternative H_1. If the null hypothesis states the absence of an effect, a Bayes factor greater than one ($\mathrm{BF}_{01} > 1$) indicates evidence for the absence of the effect and a Bayes factor smaller than one indicates evidence for the presence of the effect ($\mathrm{BF}_{01} < 1$), whereas a Bayes factor not much different from one indicates absence of evidence for either hypothesis ($\mathrm{BF}_{01} \approx 1$). A reasonable criterion for successful replication of a null result may hence be to require a Bayes factor larger than some level $\gamma > 1$ from both studies, for example, $\gamma = 3$ or $\gamma = 10$ which are conventional levels for "substantial" and "strong" evidence, respectively (***Jeffreys, 1961***). In contrast to the non-significance criterion, this criterion provides a genuine measure of evidence that can distinguish absence of evidence from evidence of absence.

The main challenge with Bayes factors is the specification of the alternative hypothesis H_1. The assumed effect under H_1 is directly related to the power of the study, and researchers who assume different effects under H_1 will end up with different Bayes factors. Instead of specifying a single effect, one therefore typically specifies a "prior distribution" of plausible effects. Importantly, the prior distribution, like the equivalence margin, should be determined by researchers with subject knowledge and before the data are collected.

To compute the Bayes factors for the RPCB null results, we used the observed effect estimates as the data and assumed a normal sampling distribution for them, as typically done in a meta-analysis. The Bayes factors BF_{01} shown in Figure 3 then quantify the evidence for the null hypothesis of no effect (H_0 : SMD $= 0$) against the alternative hypothesis that there is an effect (H_1 : SMD $\neq 0$) using a normal "unit-information" prior distribution[2] (***Kass and Wasserman, 1995***) for the effect size under the alternative H_1. We see that in most cases there is no substantial evidence for either the absence or the presence of an effect, as with the equivalence tests. For

[2]For SMD effect sizes, a normal unit-information prior is a normal distribution centered around the value of no effect with a standard deviation corresponding to one observation.

instance, with a lenient Bayes factor threshold of 3, only 1 of the 15 replications are successful, in the sense of having $\mathrm{BF}_{01} > 3$ in both the original and the replication study. The Bayes factors for the two previously discussed examples are consistent with our intuitions – in the ***Goetz et al.*** (***2011***) example there is indeed substantial evidence for the absence of an effect ($\mathrm{BF}_{01} = 5$ in the original study and $\mathrm{BF}_{01} = 4.1$ in the replication), while in the ***Dawson et al.*** (***2011***) example there is even weak evidence for the *presence* of an effect, though the Bayes factors are very close to one due to the small sample sizes ($\mathrm{BF}_{01} = 1/1.1$ in the original study and $\mathrm{BF}_{01} = 1/1.8$ in the replication).

As with the equivalence margin, the choice of the prior distribution for the SMD under the alternative H_1 is debatable. The normal unit-information prior seems to be a reasonable default choice, as it implies that small to large effects are plausible under the alternative, but other normal priors with smaller/larger standard deviations could have been considered to make the test more sensitive to smaller/larger true effect sizes. The sensitivity analysis in the appendix therefore also includes an analysis on the effect of varying prior standard deviations and the Bayes factor thresholds. However, again, to achieve replication success for a larger proportion of replications than the observed $1/15 = 7\%$, unreasonably large prior standard deviations have to be specified.

Of note, among the 15 RPCB null results, there are three interesting cases (the three effects from original paper 48) where the Bayes factor is qualitatively different from the equivalence test, revealing a fundamental difference between the two approaches. The Bayes factor is concerned with testing whether the effect is *exactly zero*, whereas the equivalence test is concerned with whether the effect is within an *interval around zero*. Due to the very large sample size in the original study ($n = 514$) and the replication ($n = 1'153$), the data are incompatible with an exactly zero effect, but compatible with effects within the equivalence range. Apart from this example, however, both approaches lead to the same qualitative conclusion – most RPCB null results are highly ambiguous.

Conclusions

We showed that in most of the RPCB studies with original "null results", neither the original nor the replication study provided conclusive evidence for the presence or absence of an effect. It seems logically questionable to declare an inconclusive replication of an inconclusive original study as a replication success. While it is important to replicate original studies with null results, our analysis highlights that they should be analyzed and interpreted appropriately. Box 1 summarizes our recommendations.

When analyzed with equivalence tests or Bayes factors, the conclusions are far less optimistic than those of the RPCB investigators, who state that if they "consider a replication to be successful overall if it succeeded on a majority of criteria, original null results were twice as likely as original positive results to mostly replicate successfully (80% vs. 40%)" (***Errington et al., 2021***, p.16).

RH: I think this is problematic because this claim relates to a definition of success as a replication succeeding on a majority of criteria, and not just the one based on significance, so I added this bit... not sure though. maybe this is also were we can say sth more about the other criteria

While the exact success rate depends on the equivalence margin and the prior distribution, sensitivity analyses showed that even with unrealistically liberal choices, the success rate remains below 40%. This is not unexpected, as a study typically requires larger sample sizes to detect the absence of an effect than to detect its presence (***Matthews, 2006***, section 11.5.3). However, the RPCB sample sizes were only chosen so that each replication had at least 80% power to detect the original effect estimate. The design of replication studies should ideally align with the planned analysis (***Anderson and Maxwell, 2017***; ***Anderson and Kelley, 2022***; ***Micheloud and Held, 2022a***; ***Pawel et al., 2022***). If the goal of the study is to find evidence for the absence of an effect, the replication sample size should also be determined so that the study has adequate power to make conclusive inferences regarding the absence of the effect.

For both the equivalence test and the Bayes factor approach, it is critical that the equivalence

Box 1: Recommendations for the analysis of replication studies of original null results. Calculations are based on effect estimates $\hat{\theta}_i$ with standard errors σ_i for $i \in \{o, r\}$ from an original study (subscript o) and its replication (subscript r). Both effect estimates are assumed to be normally distributed around the true effect size θ with known variance σ^2. The effect size θ_0 represents the value of no effect, typically $\theta_0 = 0$.

Equivalence test

1. Specify a margin $\Delta > 0$ that defines an equivalence range $[\theta_0 - \Delta, \theta_0 + \Delta]$ in which effects are considered absent for practical purposes.
2. Compute the TOST p-values for original and replication data

$$p_{\text{TOST},i} = \max\left\{\Phi\left(\frac{\hat{\theta}_i - \theta_0 - \Delta}{\sigma_i}\right), 1 - \Phi\left(\frac{\hat{\theta}_i - \theta_0 + \Delta}{\sigma_i}\right)\right\}, \; i \in \{o, r\}$$

with $\Phi(\cdot)$ the cumulative distribution function of the standard normal distribution.

```
## R function to compute TOST p-value based on effect estimate, standard error,
## null value (default is 0), equivalence margin
pTOST <- function(estimate, se, null = 0, margin) {
    p1 <- pnorm(q = (estimate - null - margin) / se)
    p2 <- 1 - pnorm(q = (estimate - null + margin) / se)
    p <- pmax(p1, p2)
    return(p)
}
```

3. Declare replication success at level α if $p_{\text{TOST},o} \leq \alpha$ and $p_{\text{TOST},r} \leq \alpha$, conventionally $\alpha = 0.05$.
4. Perform a sensitivity analysis with respect to the margin Δ. For example, visualize the TOST p-values for different margins to assess the robustness of the conclusions.

Bayes factor

1. Specify a prior distribution for the effect size θ that represents plausible values under the alternative hypothesis that there is an effect ($H_1 : \theta \neq \theta_0$). For example, specify the mean m and variance v of a normal distribution $\theta \mid H_1 \sim \text{N}(m, v)$.
2. Compute the Bayes factors contrasting $H_0 : \theta = \theta_0$ to $H_1 : \theta \neq \theta_0$ for original and replication data. Assuming a normal prior distribution, the Bayes factor is

$$\text{BF}_{01,i} = \sqrt{1 + \frac{v}{\sigma_i^2}} \exp\left[-\frac{1}{2}\left\{\frac{(\hat{\theta}_i - \theta_0)^2}{\sigma_i^2} - \frac{(\hat{\theta}_i - m)^2}{\sigma_i^2 + v}\right\}\right], \; i \in \{o, r\}.$$

```
## R function to compute Bayes factor based on effect estimate, standard error,
## null value (default is 0), prior mean (default is null value), prior variance
BF01 <- function(estimate, se, null = 0, priormean = null, priorvar) {
    bf <- sqrt(1 + priorvar/se^2) * exp(-0.5 * ((estimate - null)^2 / se^2 -
            (estimate - priormean)^2 / (se^2 + priorvar)))
    return(bf)
  }
```

3. Declare replication success at level $\gamma > 1$ if $\text{BF}_{01,o} \geq \gamma$ and $\text{BF}_{01,r} \geq \gamma$, conventionally $\gamma = 3$ (substantial evidence) or $\gamma = 10$ (strong evidence).
4. Perform a sensitivity analysis with respect to the prior distribution. For example, visualize the Bayes factors for different prior standard deviations to assess the robustness of the conclusions.

margin and the prior distribution are specified independently of the data, ideally before the original and replication studies are conducted. Typically, however, the original studies were designed

to find evidence for the presence of an effect, and the goal of replicating the "null result" was formulated only after failure to do so. It is therefore important that margins and prior distributions are motivated from historical data and/or field conventions (***Campbell and Gustafson, 2021***), and that sensitivity analyses regarding their choice are reported.

Researchers may also ask which of the two approaches is "better". We believe that this is the wrong question to ask, because both methods address slightly different questions and are better in different senses; the equivalence test is calibrated to have certain frequentist error rates, which the Bayes factor is not. The Bayes factor, on the other hand, seems to be a more natural measure of evidence as it treats the null and alternative hypotheses symmetrically and represents the factor by which rational agents should update their beliefs in light of the data. Conclusions about whether or not a study can be replicated should ideally be drawn using multiple methods. Replications that are successful with respect to all methods provide more convincing support for the original finding, while replications that are successful with only some methods require closer examination. Fortunately, the use of multiple methods is already standard practice in replication assessment (e. g., the RPCB used seven different methods), so our proposal does not require a major paradigm shift.

While the equivalence test and the Bayes factor are two principled methods for analyzing original and replication studies with null results, they are not the only possible methods for doing so. A straightforward extension would be to first synthesize the original and replication effect estimates with a meta-analysis, and then apply the equivalence and Bayes factor tests to the meta-analytic estimate. This could potentially improve the power of the tests, but consideration must be given to the threshold used for the *p*-values/Bayes factors, as naive use of the same thresholds as in the standard approaches may make the tests too liberal. Furthermore, there are various advanced methods for quantifying evidence for absent effects which could potentially improve on the more basic approaches considered here (***Lindley, 1998***; ***Johnson and Rossell, 2010***; ***Morey and Rouder, 2011***; ***Kruschke, 2018***; ***Micheloud and Held, 2022b***).

Acknowledgments

We thank the RPCB contributors for their tremendous efforts and for making their data publicly available. We thank Maya Mathur for helpful advice on data preparation. We thank Benjamin Ineichen for helpful comments on drafts of the manuscript. Our acknowledgment of these individuals does not imply their endorsement of our work. We thank the Swiss National Science Foundation for financial support (grant #189295).

Conflict of interest

We declare no conflict of interest.

Software and data

The code and data to reproduce our analyses is openly available at https://gitlab.uzh.ch/samuel.pawel/rsAbsence. A snapshot of the repository at the time of writing is available at https://doi.org/10.5281/zenodo.7906792. We used the statistical programming language R version 4.3.0 (***R Core Team, 2022***) for analyses. The R packages `ggplot2` (***Wickham, 2016***), `dplyr` (***Wickham et al., 2022***), `knitr` (***Xie, 2022***), and `reporttools` (***Rufibach, 2009***) were used for plotting, data preparation, dynamic reporting, and formatting, respectively. The data from the RPCB were obtained by downloading the files from https://github.com/mayamathur/rpcb (commit a1e0c63) and extracting the relevant variables as indicated in the R script `preprocess-rpcb-data.R` which is available in our git repository.

References

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

Anderson, S. F. and Kelley, K. (2022). Sample size planning for replication studies: The devil is in the design. *Psychological Methods*. doi:10.1037/met0000520.

Anderson, S. F. and 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.

Anderson, S. F. and Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, 52(3):305–324. doi:10.1080/00273171.2017.1289361.

Begley, C. G. and Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533. doi:10.1038/483531a.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351:1433–1436. doi:10.1126/science.aaf0918.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B., et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behavior*, 2:637–644. doi:10.1038/s41562-018-0399-z.

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

Campbell, H. and Gustafson, P. (2021). What to make of equivalence testing with a post-specified margin? *Meta-Psychology*, 5. doi:10.15626/mp.2020.2506.

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

Cooper, H., Hedges, L. V., and Valentine, J. C., editors (2019). *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation. doi:10.7758/9781610448864.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., et al. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*. doi:10.1007/s13164-018-0400-9.

Dawson, M. A., Prinjha, R. K., Dittmann, A., Giotopoulos, G., Bantscheff, M., Chan, W.-I., Robson, S. C., wa Chung, C., Hopf, C., Savitski, M. M., Huthmacher, C., Gudgin, E., Lugo, D., Beinke, S., Chapman, T. D., Roberts, E. J., Soden, P. E., Auger, K. R., Mirguet, O., Doehner, K., Delwel, R., Burnett, A. K., Jeffrey, P., Drewes, G., Lee, K., Huntly, B. J. P., and Kouzarides, T. (2011). Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. *Nature*, 478(7370):529–533. doi:10.1038/nature10509.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., and Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, 10. doi:10.7554/elife.71601.

Goetz, J. G., Minguet, S., Navarro-Lérida, I., Lazcano, J. J., Samaniego, R., Calvo, E., Tello, M., Osteso-Ibáñez, T., Pellinen, T., Echarri, A., Cerezo, A., Klein-Szanto, A. J., Garcia, R., Keely, P. J., Sánchez-Mateos, P., Cukierman, E., and Pozo, M. A. D. (2011). Biomechanical remodeling of the microenvironment by stromal caveolin-1 favors tumor invasion and metastasis. *Cell*, 146(1):148–163. doi:10.1016/j.cell.2011.05.040.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, 130(12):1005. doi:10.7326/0003-4819-130-12-199906150-00019.

Greenland, S. (2012). Nonsignificance plus high power does not imply support for the null over the alternative. *Annals of Epidemiology*, 22(5):364–368. doi:10.1016/j.annepidem.2012.02.007.

Hauck, W. W. and Anderson, S. (1986). A proposal for interpreting and reporting negative studies. *Statistics in Medicine*, 5(3):203–209. doi:10.1002/sim.4780050302.

Hoenig, J. M. and Heisey, D. M. (2001). The abuse of power. *The American Statistician*, 55(1):19–24. doi:10.1198/000313001300339897.

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon Press, third edition.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170. doi:10.1111/j.1467-9868.2009.00730.x.

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

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934. doi:10.1080/01621459.1995.10476592.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, v., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, 45:142–152. doi:10.1027/1864-9335/a000178.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, J., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490. doi:10.1177/2515245918810225.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280. doi:10.1177/2515245918771304.

Lakens, D. (2017). Equivalence tests. *Social Psychological and Personality Science*, 8(4):355–362. doi:10.1177/1948550617697177.

Lindley, D. V. (1998). Decision analysis and bioequivalence trials. *Statistical Science*, 13(2). doi:10.1214/ss/1028905932.

Makin, T. R. and de Xivry, J.-J. O. (2019). Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife*, 8. doi:10.7554/elife.48175.

Matthews, J. N. (2006). *Introduction to Randomized Controlled Clinical Trials*. Chapman and Hall/CRC, New York. doi:10.1201/9781420011302.

McCann, H. J. (2005). Intentional action and intending: Recent empirical studies. *Philosophical Psychology*, 18(6):737–748. doi:10.1080/09515080500355236.

Micheloud, C. and Held, L. (2022a). Power calculations for replication studies. *Statistical Science*, 37(3):369–379. doi:10.1214/21-sts828.

Micheloud, C. and Held, L. (2022b). The replication of non-inferiority and equivalence studies. doi:10.48550/ARXIV.2204.06960. arXiv preprint.

Morey, R. D. and Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16(4):406–419. doi:10.1037/a0024377.

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. National Academies Press. doi:10.17226/25303.

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

Pawel, S., Consonni, G., and Held, L. (2022). Bayesian approaches to designing replication studies. doi:10.48550/ARXIV.2211.02552. arXiv preprint.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10(9):712–712. doi:10.1038/nrd3439-c1.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ranganath, K. A. and Nosek, B. A. (2008). Implicit attitude generalization occurs immediately; explicit attitude generalization takes time. *Psychological Science*, 19(3):249–254. doi:10.1111/j.1467-9280.2008.02076.x.

Rufibach, K. (2009). reporttools: R functions to generate LaTeX tables of descriptive statistics. *Journal of Statistical Software, Code Snippets*, 31(1). doi:10.18637/jss.v031.c01.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680. doi:10.1007/bf01068419.

Senn, S. (2008). *Statistical Issues in Drug Development*, volume 69. John Wiley & Sons.

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

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences*, 61(8):1340–1341. doi:10.1002/jps.2600610845.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer International Publishing. doi:10.1007/978-3-319-24277-4.

Wickham, H., François, R., Henry, L., and Müller, K. (2022). *dplyr: A Grammar of Data Manipulation*. URL https://CRAN.R-project.org/package=dplyr. R package version 1.0.10.

Xie, Y. (2022). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. URL https://yihui.org/knitr/. R package version 1.40.

Appendix

Sensitivity analysis

As discussed before, the post-hoc specification of equivalence margins and the choice of the prior distribution for the SMD under the alternative H_1 is debatable and controversial. Commonly used margins in clinical research are much more stringent; for instance, in oncology, a margin of $\Delta = \log(1.3)$ is commonly used for log odds/hazard ratios, whereas in bioequivalence studies a margin of $\Delta = \log(1.25)$ is the convention (***Senn, 2008***, chapter 22). These margins would translate into margins of $\Delta = 0.14$ and $\Delta = 0.12$ on the SMD scale, respectively, using the SMD $= (\sqrt{3}/\pi) \log$ OR conversion (***Cooper et al., 2019***, p. 233). As for the Bayesian hypothesis testing we specified a normal unit-information prior under the alternative while other normal priors with smaller/larger standard deviations could have been considered. In the sensitivity analysis in the top plot of Figure 4 we show the number of successful replications as a function of the margin Δ and for different TOST *p*-value thresholds. Such an "equivalence curve" approach was first proposed by ***Hauck and Anderson*** (***1986***). We see that for realistic margins between 0 and 1, the proportion of replication successes remains below 50% for the conventional $\alpha = 0.05$ level. To achieve a success rate of $11/15 = 73\%$, as was achieved with the non-significance criterion from the RPCB, unrealistic margins of $\Delta > 2$ are required, highlighting the paucity of evidence provided by these studies. Changing the success criterion to a more lenient level ($\alpha = 0.1$) or a more stringent level ($\alpha = 0.01$) hardly changes this conclusion.

In the bottom plot of Figure 4 a sensitivity analysis is reported with respect to the choice of the prior standard deviation and the Bayes factor threshold. It is uncommon to specify prior standard deviations larger than the unit-information standard deviation of 2, as this corresponds to the assumption of very large effect sizes under the alternatives. However, to achieve replication success for a larger proportion of replications than the observed $1/15 = 7\%$, unreasonably large prior standard deviations have to be specified. For instance, a standard deviation of roughly 5 is required to achieve replication success in 50% of the replications at a lenient Bayes factor threshold of $\gamma = 3$. The standard deviation needs to be almost 20 so that the same success rate $11/15 = 73\%$ as with the non-significance criterion is achieved. The necessary standard deviations are even higher for stricter Bayes factor threshold, such as $\gamma = 6$ or $\gamma = 10$.

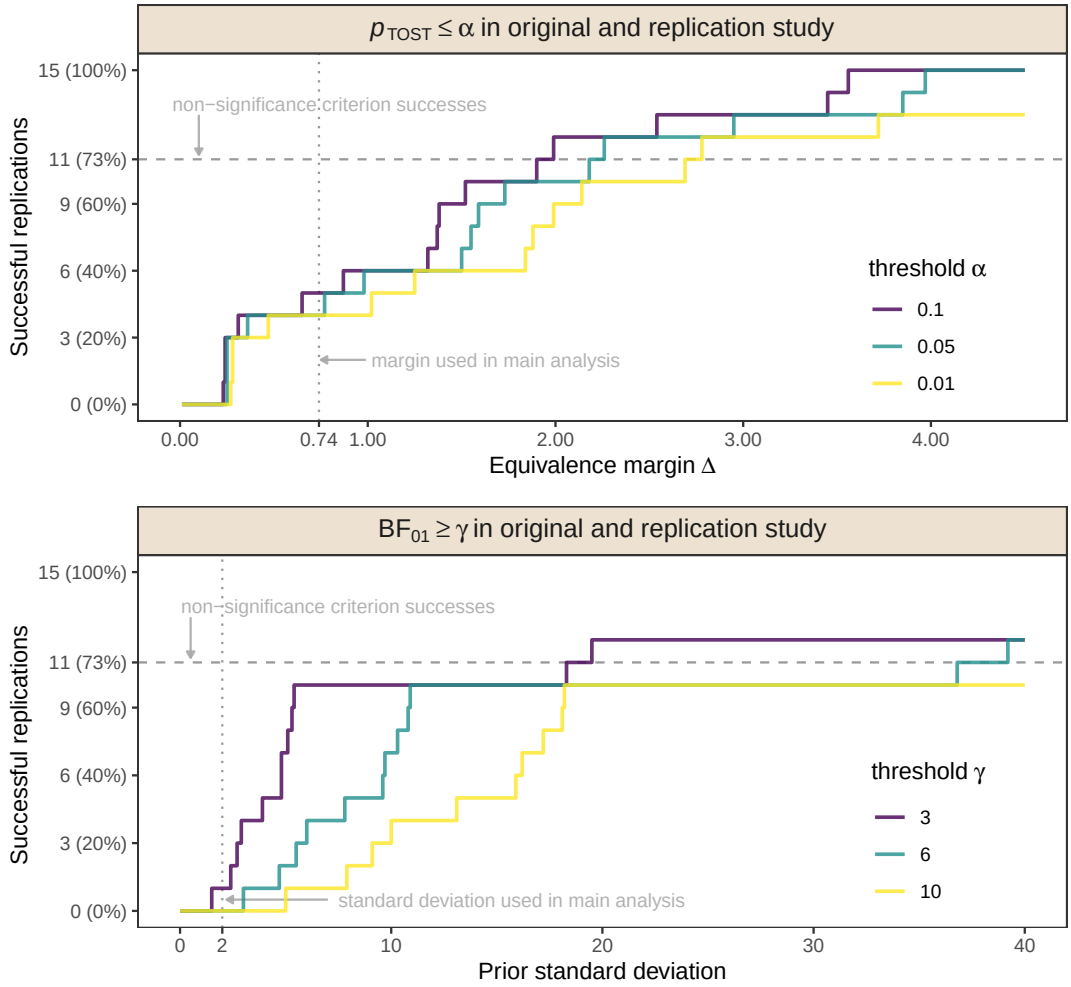

Figure 4. Number of successful replications of original null results in the RPCB as a function of the margin Δ of the equivalence test ($p_{\text{TOST}} \leq \alpha$ in both studies for $\alpha = 0.1, 0.05, 0.01$) or the standard deviation of the zero-mean normal prior distribution for the SMD effect size under the alternative H_1 of the Bayes factor test ($\text{BF}_{01} \geq \gamma$ in both studies for $\gamma = 3, 6, 10$).

Computational details

```
cat(paste(Sys.time(), Sys.timezone(), "\n"))

## 2023-06-15 09:32:35.289711 Europe/Zurich

sessionInfo()

## R version 4.3.0 (2023-04-21)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.6 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
```

```
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=de_CH.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=de_CH.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=de_CH.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=de_CH.UTF-8 LC_IDENTIFICATION=C
##
## time zone: Europe/Zurich
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] reporttools_1.1.3 xtable_1.8-4      dplyr_1.1.2       gridExtra_2.3
## [5] ggplot2_3.4.2     knitr_1.42
##
## loaded via a namespace (and not attached):
##  [1] vctrs_0.6.2       cli_3.6.1         rlang_1.1.0       xfun_0.39
##  [5] highr_0.10        generics_0.1.3    labeling_0.4.2    glue_1.6.2
##  [9] colorspace_2.1-0  scales_1.2.1      fansi_1.0.4       grid_4.3.0
## [13] munsell_0.5.0     evaluate_0.20     tibble_3.2.1      lifecycle_1.0.3
## [17] compiler_4.3.0    pkgconfig_2.0.3   farver_2.1.1      viridisLite_0.4.1
## [21] R6_2.5.1          tidyselect_1.2.0  utf8_1.2.3        pillar_1.9.0
## [25] magrittr_2.0.3    tools_4.3.0       withr_2.5.0       gtable_0.3.3
```