Design and Analysis of Replication Studies with `ReplicationSuccess`

Leonhard Held, Charlotte Micheloud, Samuel Pawel, Felix Hofmann, Florian Gerber
Epidemiology, Biostatistics and Prevention Institute (EBPI)
Center for Reproducible Science (CRS)
University of Zurich, Switzerland

Package version 1.3.3

Abstract

This vignette provides an overview of the R package `ReplicationSuccess`. The package contains utilities for the design and analysis of replication studies. Traditional methods based on statistical significance and confidence intervals, as well as the sceptical p-value (Held, 2020b) are included. The functionality of the package is illustrated using data sets from four large-scale replication projects which come also with the package.

1 Introduction

Over the course of the last decade, the conduct of replication studies has increased substantially. These developments were mainly caused by the so-called "replication crisis" in the social and life sciences. However, there is no consensus which statistical approach should be used to assess whether a replication study successfully replicated an original discovery. Moreover, depending on the chosen approach for analysis, the statistical considerations in the design of the replication study differ.

The R package `ReplicationSuccess` provides functionality to analyse and plan replication studies in several ways. Specifically, functions for analysis, power and samples size calculations based on statistical significance and confidence intervals, as well as on more recent methods, such as the sceptical p-value (Held, 2020b), are included. This vignette illustrates the usage of the package on the data sets from four large-scale replication projects which are also included in the package.

`ReplicationSuccess` was first created to provide software for computing the sceptical p-value and related power and sample size calculations (Held, 2020b). Methods to compute the p-value for intrinsic credibility (Held, 2019), the harmonic mean χ^2-test (Held, 2020a), forecasting of replication studies (Pawel and Held, 2020), and interim analyses of replication studies (Micheloud and Held, 2022) were added subsequently. Recently, substantial changes to many of the existing functions were made due to two recalibrations of the sceptical p-value approach (Held et al., 2022; Micheloud et al., 2023). Use `news(package = "ReplicationSuccess")` to see a history of the changes. The development version of `ReplicationSuccess` is available on GitHub (https://github.com/crsuzh/ReplicationSuccess/, the stable version is available on CRAN (https://cran.r-project.org/web/packages/ReplicationSuccess/).

1.1 Statistical framework

`ReplicationSuccess` assumes a simple but general statistical framework: The (suitably transformed) effect estimates $\hat{\theta}_o$ and $\hat{\theta}_r$ from original (subscript o) and replication (subscript r) study are assumed to be normally distributed around the unknown effect size θ. Their variances are equal to

their squared standard errors σ_o^2 and σ_r^2 which are assumed to be known. The same framework is also common in meta-analysis and can for example be applied to mean differences, odds ratios (log transformation), or correlation coefficients (Fisher z-transformation).

Many of the functions in the package take unitless quantities as input: the z-values $z_o = \hat{\theta}_o/\sigma_o$ and $z_r = \hat{\theta}_r/\sigma_r$, the relative effect size $d = \hat{\theta}_r/\hat{\theta}_o$ (or shrinkage $s = 1 - d$), and the variance ratio $c = \sigma_o^2/\sigma_r^2$. The squared standard errors are usually inversely proportional to the sample size of each study, $\sigma_o^2 = \kappa^2/n_o$ and $\sigma_r^2 = \kappa^2/n_r$ for some unit variance κ^2. The variance ratio can then be identified as the relative sample size $c = n_r/n_o$. This may require some adaptation for certain types of effect sizes. Computation of these quantities from real data will be illustrated below.

2 Data sets

`ReplicationSuccess` includes data from four replication projects with a "one-to-one" design (*i. e.* one replication for one original study), and from one replication project with multiple replications for one original study. The four replication projects with a one-to-one design are:

- **Reproducibility Project: Psychology:** In the *Reproducibility Project: Psychology* 100 replications of studies from the field of psychology were conducted (Open Science Collaboration, 2015). The original studies were published in three major Psychology journals in the year 2008. Only the study pairs of the "meta-analytic subset" are included here, which consists of 73 studies where the standard error of the Fisher z-transformed effect estimates can be computed (Johnson et al., 2016).

- **Experimental Economics Replication Project:** This project attempted to replicate 18 experimental economics studies published between 2011 and 2015 in two high impact economics journals (Camerer et al., 2016). For this project a *prediction market* was also conducted in order to estimate the peer beliefs about whether a replication will result in a statistically significant result. Prediction markets are a tool to aggregate beliefs of market participants regarding the possibility of an investigated outcome and they have been used successfully in numerous domains, *e. g.* sports and politics (Dreber et al., 2015). The estimated peer beliefs are also included for each study pair.

- **Social Sciences Replication Project:** This project involved 21 replications of studies on the social sciences published in the journals *Nature* and *Science* between 2010 and 2015 (Camerer et al., 2018). As in the experimental economics replication project, a prediction market to estimate peer beliefs about the replicability of the original studies was conducted and the resulting belief estimates are also provided in the package. In this project, the replications were conducted in two stages. In stage 1, the replication studies had 90% power to detect 75% of the original effect estimate. Data collection was stopped if a two-sided p-value < 0.05 and an effect in the same direction as the original were found. If not, data collection was continued in stage 2 to have 90% power to detect 50% of the original effect size for the first and second data collection pooled.

- **Experimental Philosophy Replicability Project:** In this project, 40 replications of experimental philosophy studies were carried out. The original studies had to be published between 2003 and 2015 in one of 35 journals in which experimental philosophy research is usually published (a list defined by the coordinators of this project) and they had to be listed on the experimental philosophy page of the Yale university (Cova et al., 2018). The data from the subset of 31 study pairs where effect estimates on correlation scale as well as effective sample size for both the original and replication were available are included in the package.

In these data sets, effect estimates are provided as correlation coefficients (r), as well as Fisher z-transformed correlation coefficients ($\hat{\theta} = \tanh^{-1}(r)$). In the descriptive analysis of data from replication projects it has become common practice to transform effect sizes to the correlation

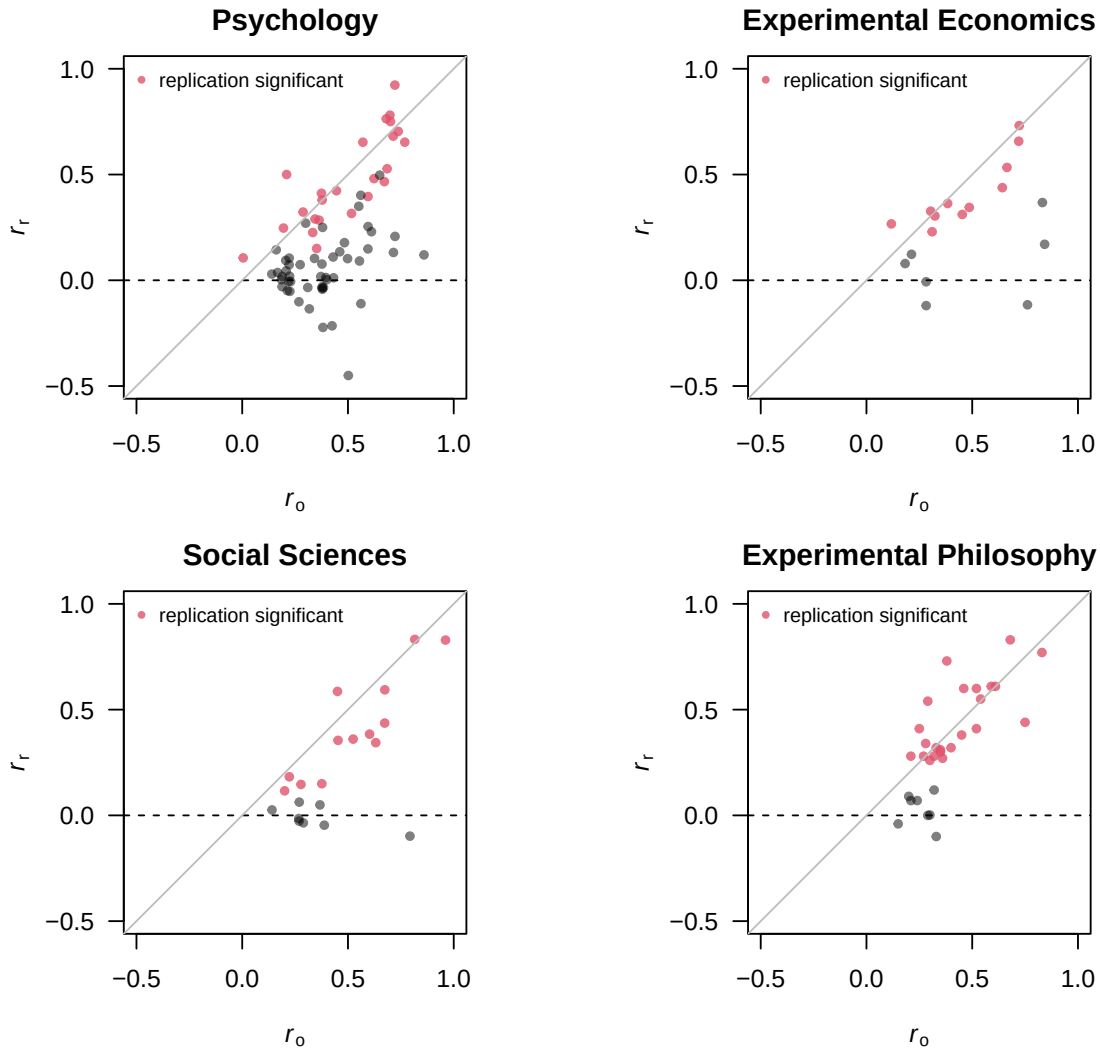
scale, because correlations are bounded to the interval between minus one and one and thus easy to compare and interpret. Design and statistical analysis, on the other hand, is then usually carried out on a scale where the distribution of the estimates is well approximated by a normal distribution. For correlation coefficients this is the case after applying the Fisher z-transformation, which leads to their variance asymptotically being only a function of the study sample size n, *i. e.* $\text{Var}(\hat{\theta}) = 1/(n-3)$ (Fisher, 1921).

The data can be loaded with the command `data("RProjects")`. For a description of the variables see the documentation with `?RProjects`. An extended version of the Social Sciences Replication Project including the details of stages one and two can be loaded with `data("SSRP")`. It is a good idea to first compute the unitless quantities z_o, z_r and c, since most functions of the package use them as input.

```
library(ReplicationSuccess)
data("RProjects")
str(RProjects, width = 80, strict.width = "cut")

## 'data.frame': 143 obs. of  15 variables:
##  $ study    : chr  "A Roelofs" "AL Morris, ML Still" "B Liefooghe, P Barrouil"..
##  $ project  : chr  "Psychology" "Psychology" "Psychology" "Psychology" ...
##  $ ro       : num  0.595 0.611 0.425 0.229 0.461 ...
##  $ rr       : num  0.14834 0.2296 -0.21524 -0.00611 0.13481 ...
##  $ fiso     : num  0.685 0.711 0.454 0.233 0.499 ...
##  $ fisr     : num  0.14944 0.23377 -0.21866 -0.00611 0.13564 ...
##  $ se_fiso  : num  0.2887 0.2132 0.2085 0.0727 0.1826 ...
##  $ se_fisr  : num  0.1925 0.2132 0.1826 0.0612 0.1474 ...
##  $ po       : num  0.017688 0.000858 0.029546 0.001368 0.006277 ...
##  $ pr       : num  0.437 0.273 0.231 0.92 0.358 ...
##  $ po1      : num  0.008844 0.000429 0.014773 0.000684 0.003139 ...
##  $ pr1      : num  0.219 0.136 0.884 0.54 0.179 ...
##  $ pm_belief: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ nr       : num  30 25 33 270 49 33 16 33 31 31 ...
##  $ no       : num  15 25 26 192 33 25 101 39 30 23 ...

## computing zo, zr, c
RProjects$zo <- with(RProjects, fiso/se_fiso)
RProjects$zr <- with(RProjects, fisr/se_fisr)
RProjects$c <- with(RProjects, se_fiso^2/se_fisr^2)
```

Note that each variable ending with an `o` is associated with the original study, while each variable ending with an `r` is associated with the replication study. Plotting the original versus the replication effect estimate on the correlation scale paints the following picture.

```
## plots of effect estimates

par(mfrow = c(2, 2), las = 1, pty = "s",
    mar = c(4, 2.5, 2.5, 1))
for (p in unique(RProjects$project)) {
    data_project <- subset(RProjects, project == p)
    significant <- ifelse(data_project$pr1 < 0.025, "#DF536BCC", "#00000080")
    plot(rr ~ ro, data = data_project, ylim = c(-0.5, 1), col = significant,
         xlim = c(-0.5, 1), main = p, xlab = expression(italic(r)[o]),
         cex = 0.7, pch = 19, ylab = expression(italic(r)[r]))
    legend("topleft", legend = "replication significant", cex = 0.8, pch = 20,
           col = "#DF536BCC", bty = "n")
    abline(h = 0, lty = 2)
    abline(a = 0, b = 1, col = "grey")
}
```



In most cases the replication estimate is smaller than the corresponding original estimate. Furthermore, a substantial number of the replication estimates does not achieve statistical significance at one-sided 2.5% level, while almost all original estimates did.

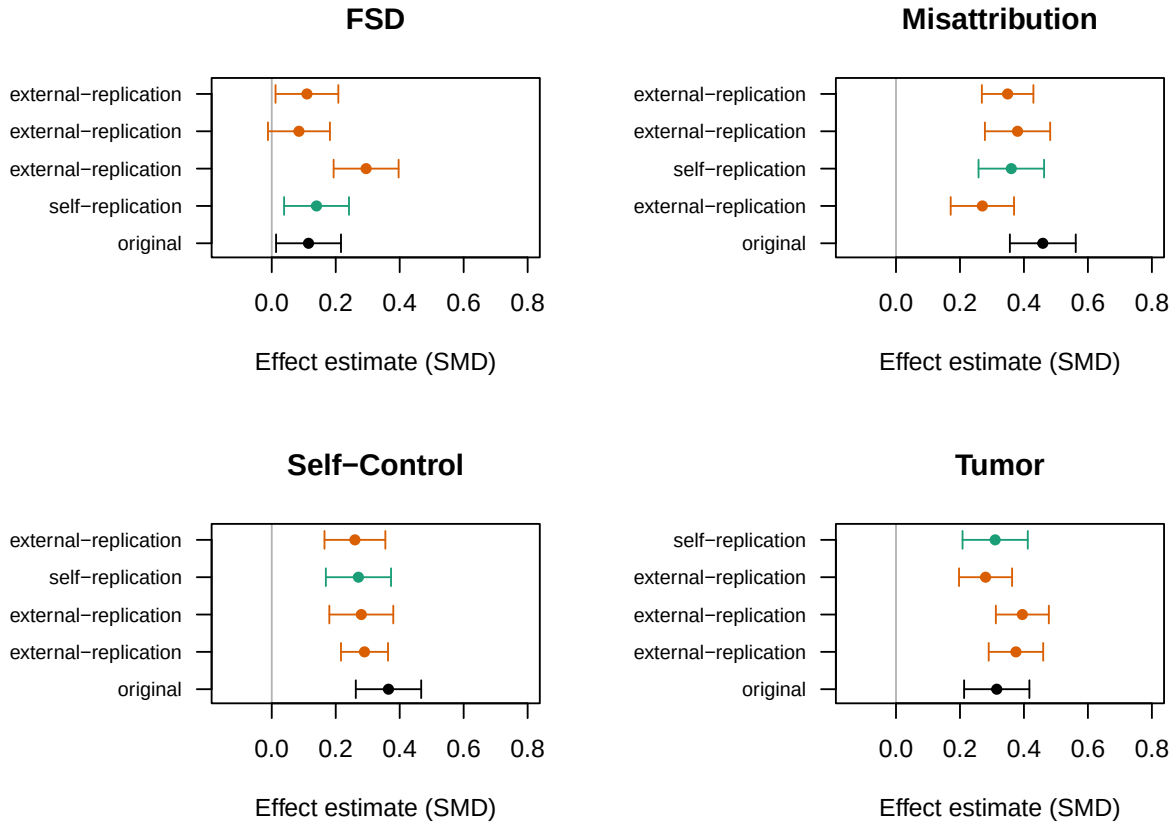
The fifth data set comes from the following project:

- **Protzko et al. (2020)**: This data set originates from a prospective replication project involving four laboratories (Protzko et al., 2020). Each of them conducted four original studies and for each original study a replication study was carried out within the same lab (self-replication) and by the other three labs (external-replication). Most studies used simple between-subject designs with two groups and a continuous outcome so that for each study, an estimate of the standardized mean difference (SMD) could be computed from the group means, group standard deviations, and group sample sizes. For studies with covariate adjustment and/or binary outcomes, effect size transformations as described in the supplementary material of Protzko et al. (2020) were used to obtain effect estimates and standard errors on SMD scale.

The dataset can be loaded with the command `data("protzko2020")`. The forest plots of the effect estimates for the first four experiments looks as follows.

```
data("protzko2020")

## forestplots of effect estimates
parOld <- par(mar = c(5, 8, 4, 2), mfrow = c(2, 2))
experiments <- unique(protzko2020$experiment)[1:4]
for (ex in experiments) {
  ## compute CIs
```

```
  dat <- subset(protzko2020, experiment == ex)
  za <- qnorm(p = 0.975)
  plotDF <- data.frame(lower = dat$smd - za*dat$se,
                       est = dat$smd,
                       upper = dat$smd + za*dat$se)
colpalette <- c("#000000", "#1B9E77", "#D95F02")
cols <- colpalette[dat$type]
yseq <- seq(1, nrow(dat))

## forestplot
plot(x = plotDF$est, y = yseq, xlim = c(-0.15, 0.8),
     ylim = c(0.8*min(yseq), 1.05*max(yseq)), type = "n",
     yaxt = "n", xlab = "Effect estimate (SMD)", ylab = "")
abline(v = 0, col = "#0000004D")
arrows(x0 = plotDF$lower, x1 = plotDF$upper, y0 = yseq, angle = 90,
       code = 3, length = 0.05, col = cols)
points(y = yseq, x = plotDF$est, pch = 20, lwd = 2, col = cols)
axis(side = 2, at = yseq, las = 1, labels = dat$type, cex.axis = 0.85)
title(main = ex)
}
```



3 Design and analysis of replication studies

Although a replication study needs to be planned and conducted before the results can be analysed, we will first discuss the particular analysis approaches. We do this because the chosen analysis strategy has a substantial impact on the design of a replication study. In the design phase of a replication study, we will then focus only on the determination of the sample size.

3.1 Statistical significance

Analysis The most commonly used approach is to declare a replication successful if both the original and replication study achieve statistical significance (in the same direction). The significance level is conventionally chosen to be 0.05 for two-sided p-values, respectively, 0.025 for one-sided p-values. Working with one-sided p-values is usually simpler since effect direction is automatically taken into account, while with two-sided p-values one has also to check whether the original and replication estimate go in the same direction. For the four one-to-one replication projects we can simply check whether the one-sided p-values (in the positive direction) of original and replication are both below the conventional threshold 0.025.

```
for (p in unique(RProjects$project)) {
    data_project <- subset(RProjects, project == p)
    significant_O <- data_project$po1 < 0.025
    significant_R <- data_project$pr1 < 0.025
    success <- significant_O & significant_R
    cat(paste0(p, ": \n"))
    cat(paste0(round(mean(significant_O)*100, 1), "% original studies significant (",
               sum(significant_O), "/", length(significant_O), ")\n"))
    cat(paste0(round(mean(significant_R)*100, 1), "% replications significant (",
               sum(significant_R), "/", length(significant_R), ")\n"))
    cat(paste0(round(mean(success)*100, 1),
               "% both studies significant in the same direction (",
               sum(success), "/", length(success), ")\n \n"))
}

## Psychology:
## 89% original studies significant (65/73)
## 32.9% replications significant (24/73)
## 28.8% both studies significant in the same direction (21/73)
##
## Experimental Economics:
## 88.9% original studies significant (16/18)
## 61.1% replications significant (11/18)
## 55.6% both studies significant in the same direction (10/18)
##
## Social Sciences:
## 100% original studies significant (21/21)
## 61.9% replications significant (13/21)
## 61.9% both studies significant in the same direction (13/21)
##
## Experimental Philosophy:
## 96.8% original studies significant (30/31)
## 74.2% replications significant (23/31)
## 74.2% both studies significant in the same direction (23/31)
##
```

Despite its appealing simplicity, assessing replication success with statistical significance is often criticized. For example, non-significant replication results are expected if the original finding was a false positive, however, they can also be caused due to low power of the replication study (Goodman, 1992). On the other hand, statistical significance can still be achieved for a replication effect estimate which is much smaller than the one from the original study, provided its standard error is small enough (*e. g.* because of a very large replication sample size).

Design Selecting the same sample size in the replication study as in the original study may lead to a severely underpowered design and as a result, true effects may not be detected. To ensure that the replication study reliably detects true effects, the studies should be well-powered. In classical sample size planning, usually a clinically relevant effect is specified and the sample size is then

determined so that it can be detected with a certain power. Luckily, in the replication setting the clinically relevant effect does not need to be specified but can be replaced with the effect estimate from the original study. However, using the standard sample size calculation approach is not well suited, because the uncertainty of the original effect estimate is ignored.

One way of tackling this issue is to use a Bayesian approach, incorporating the original estimate and its precision into a design prior that is used for power calculations. This corresponds to the concept of "predictive power" and generally leads to larger sample sizes than the standard method. In practice, however, often more ad hoc approaches are used. For instance, the original estimate is just shrunken by an (arbitrary) constant, *e. g.* it was halved in the social sciences replication project, and standard sample size calculations are then carried out.

Using the function `sampleSizeSignificance`, it is straightforward to plan the sample size of the replication study with the just mentioned approaches. The argument `designPrior` allows to carry out sample size planning based on classical power ignoring the uncertainty (`"conditional"`) or based on predictive power (`"predictive"`). Moreover, ad hoc shrinkage can be specified with the argument `shrinkage`. Note that the function `sampleSizeSignificance`, as well as most of the functions from the package, takes z-values (and not p-values) as arguments. The transformation from p- to z-values and vice versa can easily be done using the functions `p2z` and `z2p`.

The following code shows a few examples. Note that the function returns the required relative sample size $c = n_r/n_o$, *i. e.* the factor by which the sample size of the original study needs to be multiply for the replication study.

```
sampleSizeSignificance(zo = 2.5, power = 0.8, level = 0.05, designPrior = "conditional")

## [1] 0.9892092

sampleSizeSignificance(zo = 2.5, power = 0.8, level = 0.05, designPrior = "predictive")

## [1] 1.388112

sampleSizeSignificance(zo = 2.5, power = 0.8, level = 0.05, designPrior = "conditional",
                       shrinkage = 0.25)

## [1] 1.758594
```

The power of the replication study for a fixed relative sample size c can be calculated with the function `powerSignificance`. Figure 1 shows the power to achieve significance in the replication as a function of either the (one-sided) p-value or the z-value of the original study. If the original estimate was just significant at the 0.025 level, the probability for significance in the replication is just about 0.5 for conditional and predictive power. This result was first mentioned by Goodman (1992) already two decades ago, yet many practitioners of statistics still find it counterintuitive, because they confuse type I error rates with replication probabilities. Thus, for the replication to achieve significance with high probability, the sample size needs to be increased compared to the original if the the evidence for the original discovery was only weak or moderate (Figure 2).

3.2 Compatibility of effect size

Analysis Another way for analysing a replication study is to examine the statistical compatibility between original and replication effect estimate. A popular approach is to check whether the replication estimate is contained within a prediction interval based on the original estimate (Patil et al., 2016; Pawel and Held, 2020). This approach based on a $(1 - \alpha)$ prediction interval is equivalent to conducting a meta-analytic Q-test with the two estimates, and rejecting compatibility when the corresponding p-value $p_Q < \alpha$ (Hedges and Schauer, 2019). The p-value from the Q-test is usually preferred, since it tells quantitatively how compatible the estimates are. In contrast, a prediction interval can give a better idea about the range of plausible replication effect estimates besides the observed one. Both approaches are available in `ReplicationSuccess`: The function `Qtest` returns

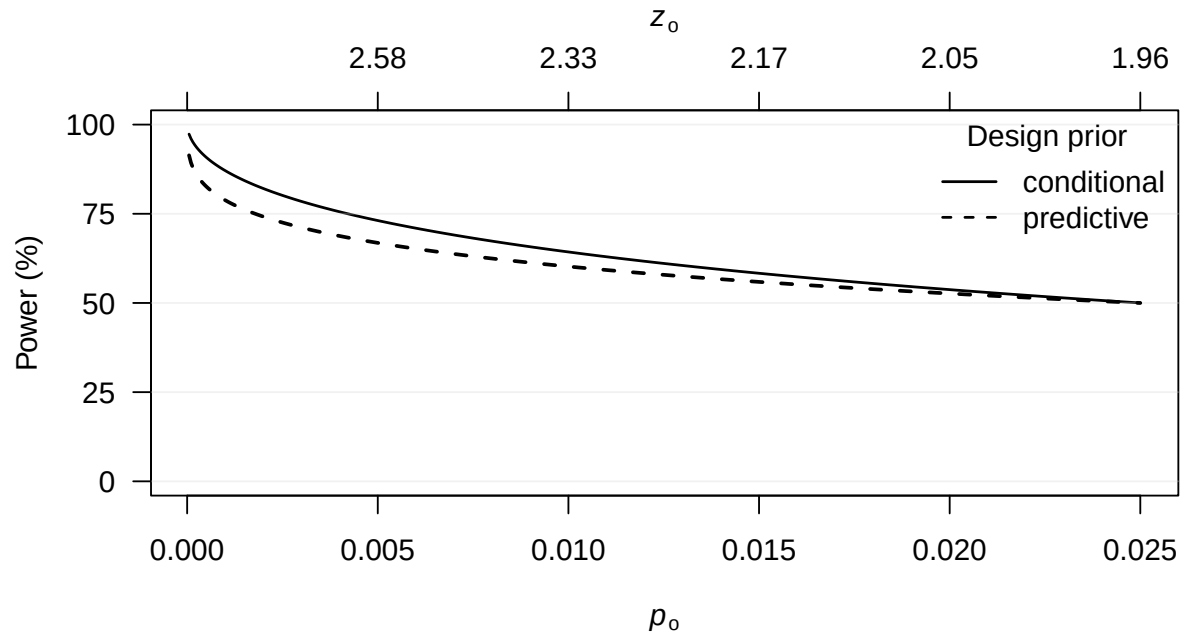


Figure 1: Power to achieve significance of the replication study at the one-sided 2.5% level as a function of the (one-sided) p-value or z-value of the original study using the same sample size as in the original study.

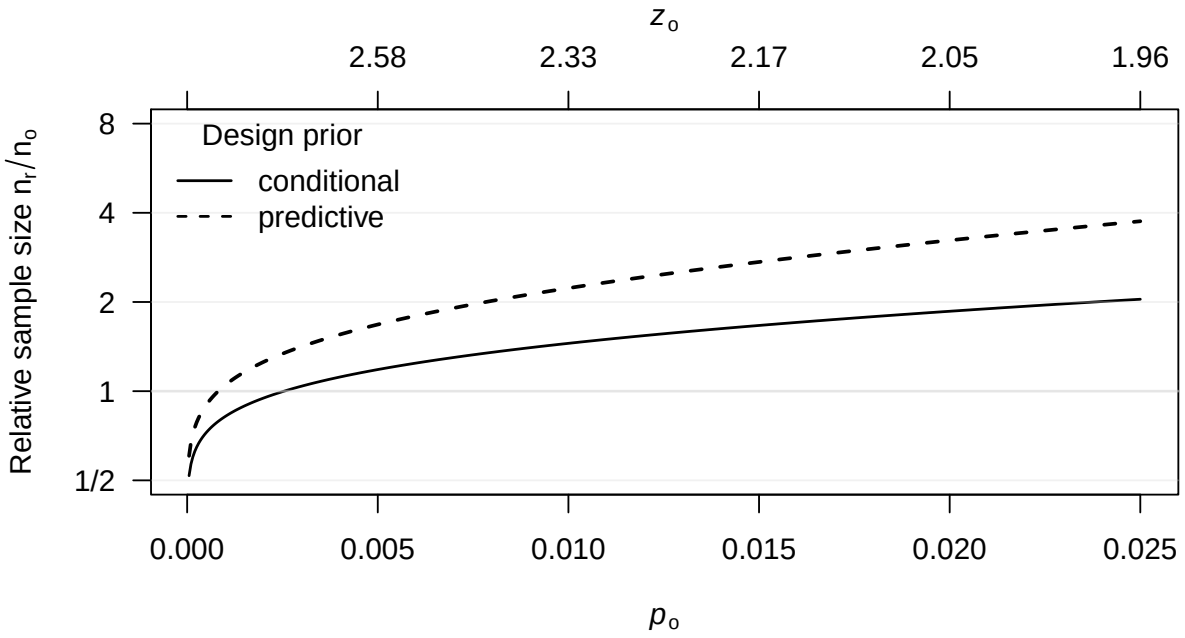


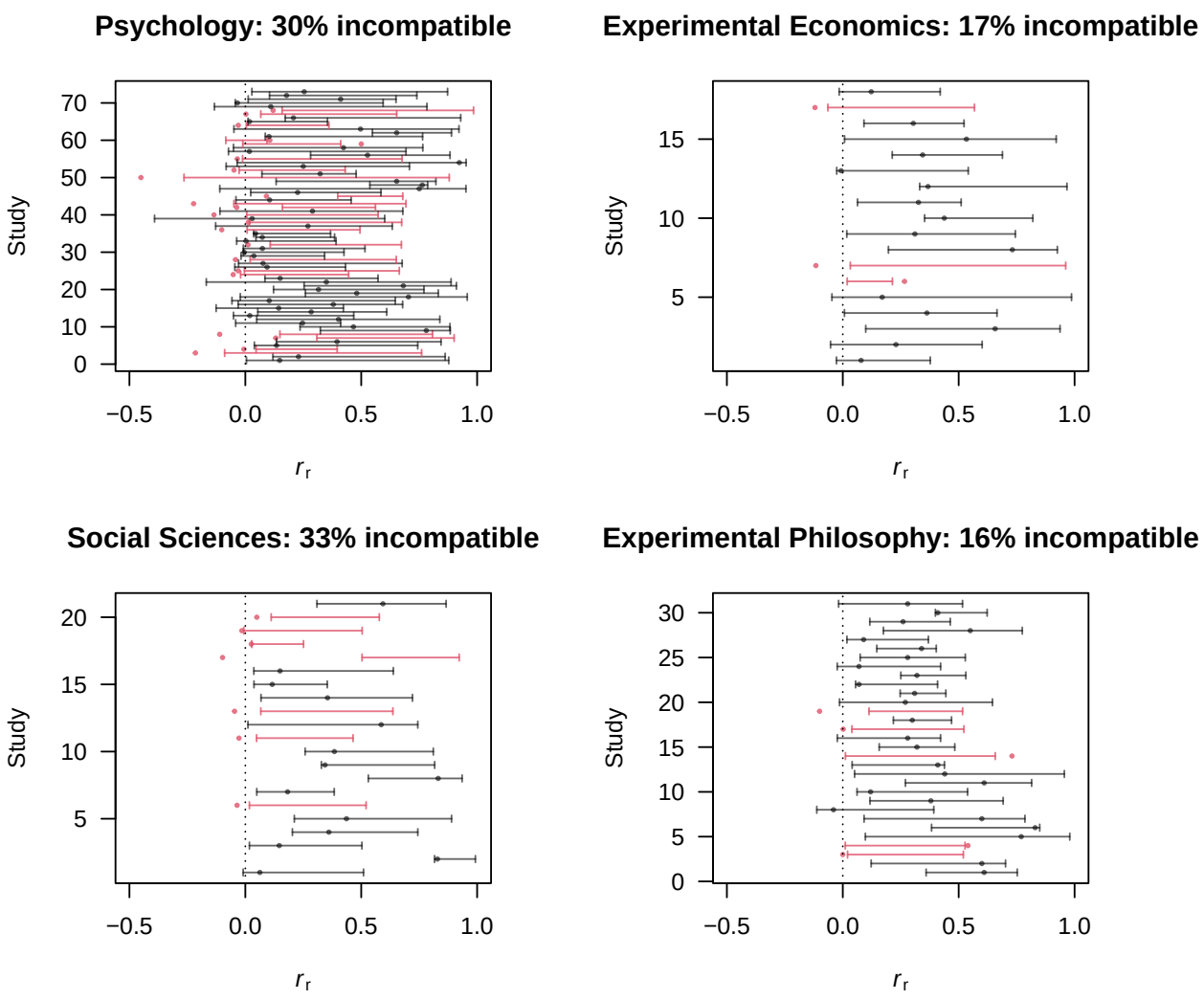
Figure 2: Relative sample size to achieve significance of the replication study at the one-sided 2.5% level with 80% power as a function of the (one-sided) p-value or z-value of original study.

the p-value from the meta-analytic Q-test, whereas the function `predictionInterval` returns a prediction interval for the replication effect based on the original counterpart (see the documentation of the functions for further details).

For the four data sets, we can easily compute p-values from Q-test, as well as 95% prediction in-

tervals. For easier visual assessment we transform the intervals and estimates back to the correlation scale.

```
## compute prediction intervals for replication projects
par(mfrow = c(2, 2), las = 1, mai = rep(0.65, 4))
for (p in unique(RProjects$project)) {
    data_project <- subset(RProjects, project == p)
    pQ <- Qtest(thetao = data_project$fiso,
                thetar = data_project$fisr,
                seo = data_project$se_fiso,
                ser = data_project$se_fisr)
    PI <- predictionInterval(thetao = data_project$fiso,
                             seo = data_project$se_fiso,
                             ser = data_project$se_fisr)
    ## transforming back to correlation scale
    PI <- tanh(PI)
    incompatible <- pQ < 0.05 ## incompatible at 5% level
    color <- ifelse(incompatible == FALSE, "#00000099", "#DF536BCC")
    study <- seq(1, nrow(data_project))
    plot(data_project$rr, study, col = color, pch = 20, cex = 0.5,
         xlim = c(-0.5, 1), xlab = expression(italic(r)[r]), ylab = "Study",
         main = paste0(p, ": ", round(mean(incompatible)*100, 0), "% incompatible"))
    arrows(PI$lower, study, PI$upper, study, length = 0.02,
           angle = 90, code = 3, col = color)
    abline(v = 0, lty = 3)
}
```



While both approaches enable statements about compatibility of original and replication effect estimates, they carry a fundamental problem due to the structure of the underlying hypothesis test: If $p_Q < \alpha$ (or equivalently the $(1-\alpha)$ prediction interval does not contain the replication estimate) one has established incompatibility/non-replication. If, however, one fails to establish incompatibility, it remains unclear whether this is due to small power or true compatibility of both estimates (Hedges and Schauer, 2019).

3.3 The sceptical p-value

Analysis The *sceptical p-value*, a new quantitative measure of replication success was first introduced in Held (2020b). Conceptually, replication success is declared if the replication study is in conflict with a sceptical prior that would render the original study non-significant. The sceptical p-value arises from combining the analysis of credibility (Matthews, 2001) with the prior-predictive check (Box, 1980). Specifically, using Bayes theorem in reverse, the prior distribution of the effect size is determined such that based on the original study, the $(1-\alpha)$ credible interval of the posterior distribution of the effect just includes zero. This prior corresponds to the objection of a sceptic who argues that the original finding is no longer significant if combined with a sufficiently sceptical prior. Replication success at level α is then achieved if the tail probability of the replication estimate under its prior predictive distribution is smaller than α, rendering the objection of the sceptic unrealistic. The smallest level α at which replication success can be declared defines the sceptical p-value.

The method has attractive properties: The sceptical p-value is never smaller than the ordinary p-values from both studies which ensures that both studies have to be sufficiently convincing on their own such that replication success is possible. It also takes into account the size of the effect estimates, *i. e.* it becomes larger if the replication estimate is smaller than the original estimate, which guarantees that shrinkage of the replication effect estimate is penalized. One-sided sceptical p-values are preferred because they guarantee that replication success is only possible if the directions of original and replication effect estimates are the same. We will only report one-sided p-values in the following. The sceptical p-value in its original formulation ('nominal') can be easily computed with the function `pSceptical` and `type = "nominal"`.

Figure 3 shows the original study from Morewedge et al. (2010) and its replication by the SSRP (Camerer et al., 2018). In this example, the one-sided sceptical p-value turns out to be $p_S = 0.036$.

The example study can be isolated with the following command:

```
morewedge <- subset(RProjects, study == "Morewedge et al. (2010), Science")
```

Interpretation and recalibration

Interpretation of the sceptical p-value as a continuous measure of replication success is recommended. However, if an answer to "did it replicate?" is needed, a threshold is required. The sceptical p-value can be thresholded at α just as an ordinary p-value. Our default choice is $\alpha = 0.025$ for one-sided p-values.

Held (2020b) calculated the nominal sceptical p-values of a number of case studies and a relatively low rate of replication success was found. In addition, Held et al. (2022) further studied properties of the nominal sceptical p-value and concluded that it may be too stringent for most realistic scenarios. Two recalibrations were subsequently proposed: the golden and the controlled sceptical p-values. They can be computed using the `pSceptical` function with `type = "golden"` or `type = "controlled"`, respectively, and should also be thresholded at α.

The two recalibration types can be summarized as follows:

- The *golden* sceptical p-value (Held et al., 2022) provides an attractive balance between significance testing and effect size comparison. It ensures that for original studies, which are just significant at the specified significance level α, replication success is only possible if the replication effect estimate is larger than the original one (*i. e.* if there is no shrinkage). The golden sceptical p-value controls the overall Type-I error rate for $c \geq 1$, so does not provide exact

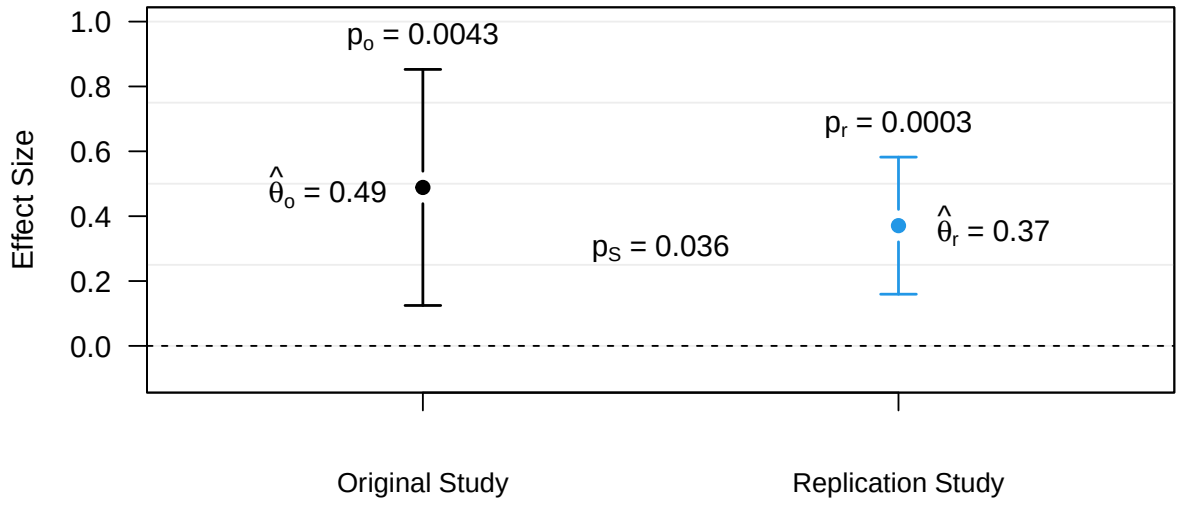


Figure 3: Original study from Morewedge et al. (2010) and its replication by the SSRP (Camerer et al., 2018). Shown are the effect estimates $\hat{\theta}_o$, $\hat{\theta}_r$ together with the 95% confidence interval and the one-sided p-values p_o and p_r and the nominal sceptical p-value p_S.

overall Type-I error control for any c. This is our recommended default type of recalibration. In the Morewedge et al. (2010) replication example, we obtain $p_S = 0.011 < 0.025$, so the replication is successful with the golden sceptical p-value.

- The *controlled* sceptical p-value (Micheloud et al., 2023) ensures exact overall Type-I error control for any value of the variance ratio c. The overall T1E rate is controlled at level α^2 for `alternative = "two.sided"` (where α is two-sided) or `"one.sided"` (where α is one-sided) if the direction was pre-specified in advance. In the Morewedge et al. (2010) replication example, we obtain $p_S = 0.0026 < 0.025$, so the replication is also successful with the controlled sceptical p-value.

The three types of sceptical p-values for the Morewedge et al. (2010) example are calculated as follows:

```
print(pS_nominal <- pSceptical(zo = morewedge$zo, zr = morewedge$zr,
                           c = morewedge$c,  alternative = "one.sided",
                           type = "nominal"))

## [1] 0.03559125

print(pS_golden <- pSceptical(zo = morewedge$zo, zr = morewedge$zr,
                          c = morewedge$c,  alternative = "one.sided",
                          type = "golden"))

## [1] 0.01086314

print(pS_controlled <- pSceptical(zo = morewedge$zo, zr = morewedge$zr,
                              c = morewedge$c,  alternative = "one.sided",
                              type = "controlled"))

## [1] 0.002619875
```

The one-sided golden and controlled sceptical p-values are computed for the four one-to-one replication projects as follows:

```
## computing one-sided golden and controlled sceptical p-value for replication projects
RProjects$psG <- with(RProjects,
                      pSceptical(zo = zo, zr = zr, c = c,
                                 alternative = "one.sided", type = "golden"))
RProjects$psC <- with(RProjects,
                      pSceptical(zo = zo, zr = zr, c = c,
                                 alternative = "one.sided", type = "controlled"))
```

And the success rates in the four projects with the statistical significance criterion, and the golden and controlled sceptical p-value are

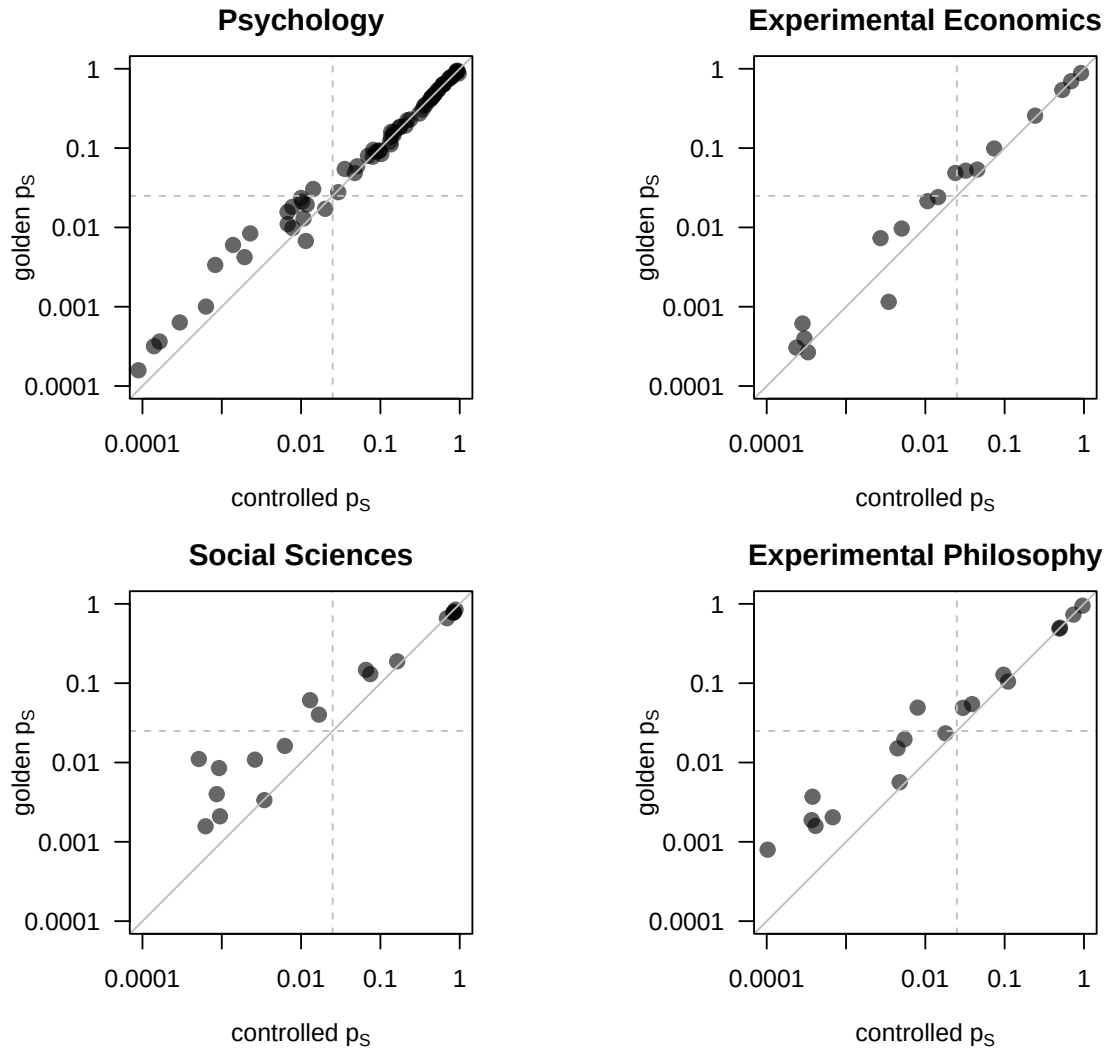
```
## Psychology:
## 30.14% smaller than 0.025 (one-sided golden sceptical p-value)
## 31.51% smaller than 0.025 (one-sided controlled sceptical p-value)
## 28.77% smaller than 0.025 (both one-sided traditional p-values)
## Discrepant studies (golden vs significance):
##    ro   rr   c     po1      pr1   psG
##  0.20 0.25 2.6 0.02800 0.000047 0.024
##  0.56 0.40 0.6 0.00026 0.035000 0.017
##  0.35 0.15 2.7 0.00140 0.023000 0.031
## Discrepant studies (controlled vs significance):
##    ro   rr   c     po1      pr1  psC
##  0.20 0.25 2.6 0.02800 0.000047 0.01
##  0.56 0.40 0.6 0.00026 0.035000 0.02
## Discrepant studies (golden vs controlled):
##    ro   rr   c    po1   pr1   psG   psC
##  0.35 0.15 2.7 0.0014 0.023 0.031 0.014
##
##
## Experimental Economics:
## 55.56% smaller than 0.025 (one-sided golden sceptical p-value)
## 61.11% smaller than 0.025 (one-sided controlled sceptical p-value)
## 55.56% smaller than 0.025 (both one-sided traditional p-values)
## Discrepant studies (controlled vs significance):
##    ro   rr   c   po1    pr1   psC
##  0.31 0.23 3.2 0.027 0.0059 0.024
## Discrepant studies (golden vs controlled):
##    ro   rr   c   po1    pr1   psG   psC
##  0.31 0.23 3.2 0.027 0.0059 0.049 0.024
##
##
## Social Sciences:
## 52.38% smaller than 0.025 (one-sided golden sceptical p-value)
## 61.9% smaller than 0.025 (one-sided controlled sceptical p-value)
## 61.9% smaller than 0.025 (both one-sided traditional p-values)
## Discrepant studies (golden vs significance):
##    ro   rr   c    po1    pr1   psG
##  0.28 0.15 3.5 0.0089 0.0110 0.040
##  0.38 0.15 9.2 0.0110 0.0043 0.061
## Discrepant studies (golden vs controlled):
##    ro   rr   c    po1    pr1   psG   psC
##  0.28 0.15 3.5 0.0089 0.0110 0.040 0.017
##  0.38 0.15 9.2 0.0110 0.0043 0.061 0.013
##
##
## Experimental Philosophy:
## 70.97% smaller than 0.025 (one-sided golden sceptical p-value)
```

```
## 74.19% smaller than 0.025 (one-sided controlled sceptical p-value)
## 74.19% smaller than 0.025 (both one-sided traditional p-values)
## Discrepant studies (golden vs significance):
##    ro   rr   c   po1    pr1   psG
##  0.75 0.44 9.4 0.015 0.0006 0.049
## Discrepant studies (golden vs controlled):
##    ro   rr   c   po1    pr1   psG   psC
##  0.75 0.44 9.4 0.015 0.0006 0.049 0.008
##
##
```

For most studies the three approaches agree, but there is disagreement for seven of them. The golden sceptical p-value may not indicate replication success when there is substantial shrinkage of the replication effect estimate relative to the original one, even if both estimates are significant (this is the case for one study in the psychology project, two studies in the social sciences project, one study in the philosophy project). In contrast, provided there is not much shrinkage, it may still indicate replication success for non-significant original or replication studies (this is the case for two studies in the psychology project). On the other hand, the controlled sceptical p-value can indicate success for non-significant original or replication study, even in the presence of considerable shrinkage (this is the case for one study in the psychology project and one in the experimental economics project).

The following plot shows golden p_S versus controlled p_S for all study pairs, stratified by project.

```
par(mfrow = c(2, 2), las = 1, pty = "s",
    mar = c(4, 2.5, 2.5, 1))
  myaxis <- c(10e-5, 0.001, 0.01, 0.1, 1)
for (p in unique(RProjects$project)) {
  data_project <- subset(RProjects, project == p)
  plot(psG ~ psC, data = data_project, ylim = c(10e-5, 1),
       xlim = c(10e-5, 1), main = p, xlab = expression(paste("controlled ", p[S])),
       ylab = expression(paste("golden ", p[S])),
       pch = 19,
       col = "#00000099",
       cex = 1.3,
       axes = F,
       log = "xy")
  abline(h = 0, lty = 2)
  abline(a = 0, b = 1, col = "grey")
  abline(h = 0.025, lty = 2, col = "grey")
  abline(v = 0.025, lty = 2, col = "grey")

axis(1, at = myaxis, labels = myaxis)
axis(2,at = myaxis, labels = myaxis)
box()
}
```



Replication success level

Instead of recalibrating the sceptical p-value and thresholding it at α, it is equivalent to use the uncalibrated sceptical p-value and threshold it at the replication success level γ. The replication success level is computed using `levelSceptical` and the recalibration `type` (`"golden"` or `"controlled"`). The golden replication success level depends on α and is 0.062 for $\alpha = 0.025$. The controlled replication success level depends on α and the variance ratio c. It is for example 0.065 for $c = 1$ and 0.083 for $c = 2$. Using `type = "nominal"` simply returns α.

```
## computing nominal, golden and controlled replication success levels
## for one-sided uncalibrated sceptical p-value

print(rs_level_nom <- levelSceptical(level = 0.025, alternative = "one.sided",
                                     type = "nominal"))

## [1] 0.025

print(rs_level_gol <- levelSceptical(level = 0.025, alternative = "one.sided",
                                     type = "golden"))

## [1] 0.06167928

print(rs_level_contr <- levelSceptical(level = 0.025, alternative = "one.sided",
                                       type = "controlled", c = c(1, 2)))

## [1] 0.06530978 0.08296757
```

The replication success level γ is also the bound on the partial Type-I error rate of the sceptical p-value (Micheloud et al., 2023, Section 3.1). This means that the individual p-values p_o and p_r both need to be smaller than γ for replication success to be possible with the sceptical p-value.

Design Sample size calculations work in a similar manner as when they are based on statistical significance: Using the function `sampleSizeReplicationSuccess`, we need to choose a design prior, a threshold α for the sceptical p-value, a recalibration type, and the desired power to obtain the required relative sample size $c = n_r/n_o$. The following code shows two examples.

```
sampleSizeReplicationSuccess(zo = 2.5, power = 0.8, level = 0.025,
                             alternative = "one.sided",
                             designPrior = "conditional",
                             type = c("golden", "controlled"))

## [1] 1.377053 1.201714

sampleSizeReplicationSuccess(zo = 2.5, power = 0.8, level = 0.025,
                             alternative = "one.sided",
                             designPrior = "predictive",
                             type = c("golden", "controlled"))

## [1] 2.784714 1.735828
```

The function `powerSignificance` allows to calculate the power for a fixed relative sample size c. Figure 4 shows the power to achieve replication success with the golden and controlled sceptical p-values as a function of the one-sided p-value (or z-value) of the original study, assuming equal sample sizes in original and replication studies ($c = 1$).

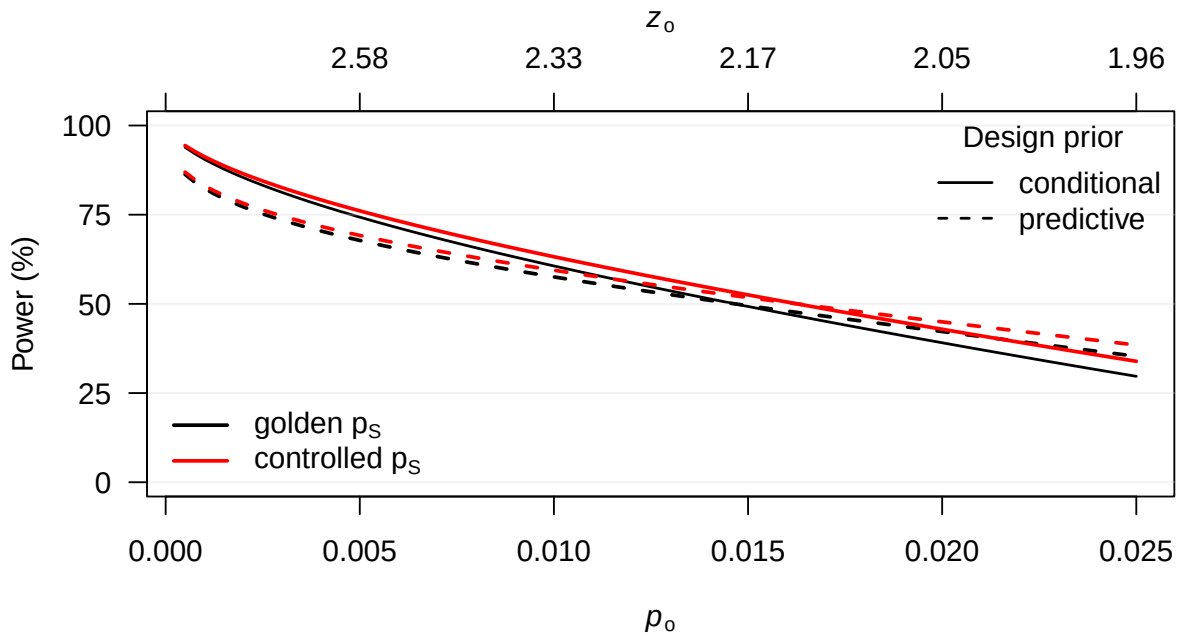


Figure 4: Power to achieve replication success (with the golden and controlled sceptical p-value and $\alpha = 0.025$) as a function of the one-sided p-value or z-value of the original study.

Figure 5 shows the required sample size to achieve replication success with the golden and controlled sceptical p-value with 80% power. With the golden sceptical p-value, the required relative sample sizes gets very large for borderline significant original studies and is even impossible for non-significant original studies.

We thus recommend to use the controlled sceptical p-value for sample size calculations, which also allows sample size calculation for non-significant original studies.

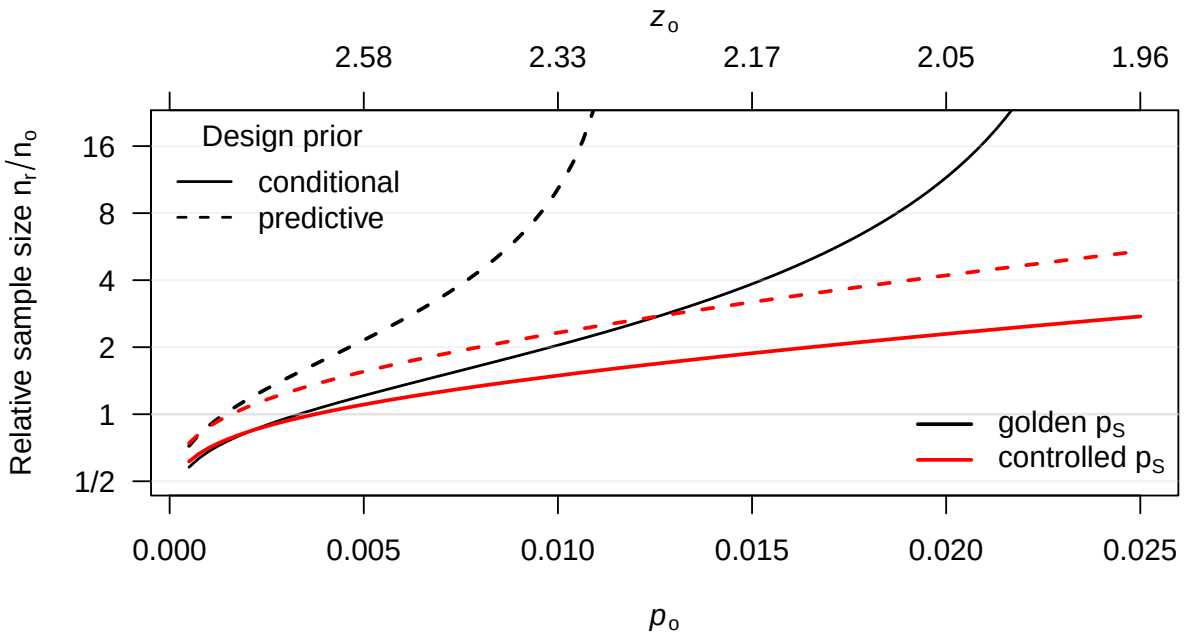


Figure 5: Relative sample size to achieve replication success (with the golden and controlled sceptical p-values and $\alpha = 0.025$) with 80% power as a function of the (one-sided) p-value or z-value of the original study.

3.4 Relative effect size

Analysis The requirements on the replication p-value ($p_r < \alpha$) and on the sceptical p-value ($p_S < \alpha$) can both be transformed to requirements on the relative effect size $d = \hat{\theta}_r/\hat{\theta}_o$ (Held et al., 2022). In short, replication success is declared if the relative effect size is larger than a certain bound (the minimum relative effect size $d_{\min}$). The minimum relative effect size $d_{\min}$ is computed based on the result from the original study, the relative sample size, the significance level/threshold for the sceptical p-value. The type of recalibration is also required for the sceptical p-value. The functions are `effectSizeSignificance` and `effectSizeReplicationSuccess` for significance and the sceptical p-value, respectively.

For the Morewedge et al. (2010) example, the minimum relative effect size to achieve replication success is $d_{\min} = 0.43$ for significance and $d_{\min} = 0.54$ with the golden sceptical p-value. Since the observed relative effect size is $d = 0.76$, the replication is successful with both methods.

4 Special topics

4.1 Interim analysis

Adaptive designs are a type of designs where one or more interim analyses are planned during the course of a study. This topic has extensively been studied and used in clinical trials for example, where continuing a study that should be stopped may lead to serious consequences. However, this type of design has rarely been covered in the framework of replication studies. An adaptive design was adopted in the social sciences replication project, but without a power (re)calculation at interim.

`ReplicationSuccess` allows to calculate the power of the replication study after an interim analysis has been performed, taking into account the results from the first part of the study. The

power at interim is a useful tool to decide whether a replication study should be stopped prematurely for futility (Micheloud and Held, 2022). The function `powerSignificanceInterim` is an extension of `powerSignificance` and requires in addition the specification of `zi`, the z-value at the interim analysis and `f`, the fraction of the replication study already completed. Moreover, the argument `designPrior` can be set to `"conditional"`, `"informed predictive"` and `"predictive"`. Finally, the argument `analysisPrior` allows to also take the original result into account in the analysis of the replication study.

4.2 Between-study heterogeneity

It is often more realistic to assume that the unknown effect sizes from original and replication studies are not exactly the same but that there is between-study heterogeneity. This can be the case, for example, if the replication study is conducted with a different population of participants (*e. g.* in a different country) or in another laboratory with different equipment. For this reason, the functions for design of replication studies allow to incorporate additionally uncertainty due to between-study heterogeneity. Some functions in the package (*e. g.* `sampleSizeSignificance` or `predictionInterval`) allows to specify the argument `h`, the relative between-study heterogeneity variance $h = \tau^2/\sigma^2$, *i. e.* the ratio of the heterogeneity variance to the variance of the original effect estimate. By default, `h` is set to zero, however, if between-study heterogeneity is expected, *e. g.* a different population of study participants is used, this should be considered in the design. See Pawel and Held (2020) for more details.

4.3 Data-driven shrinkage with empirical Bayes

The functions for design of replication studies allow to specify the argument `shrinkage`, in order to shrink the original effect estimate towards zero by a certain (arbitrary) amount. A more principled approach is to use a design prior which induces shrinkage and then estimate the prior variance by empirical Bayes. This leads to "data-driven" shrinkage that is larger when there was only weak evidence for the effect, and smaller when there was strong evidence for the effect (shown in Figure 6). Furthermore, under this prior, the specified between-study heterogeneity will also induce shrinkage towards zero, for details see Pawel and Held (2020). Empirical Bayes shrinkage can be enabled by setting the design prior argument to `"EB"`.

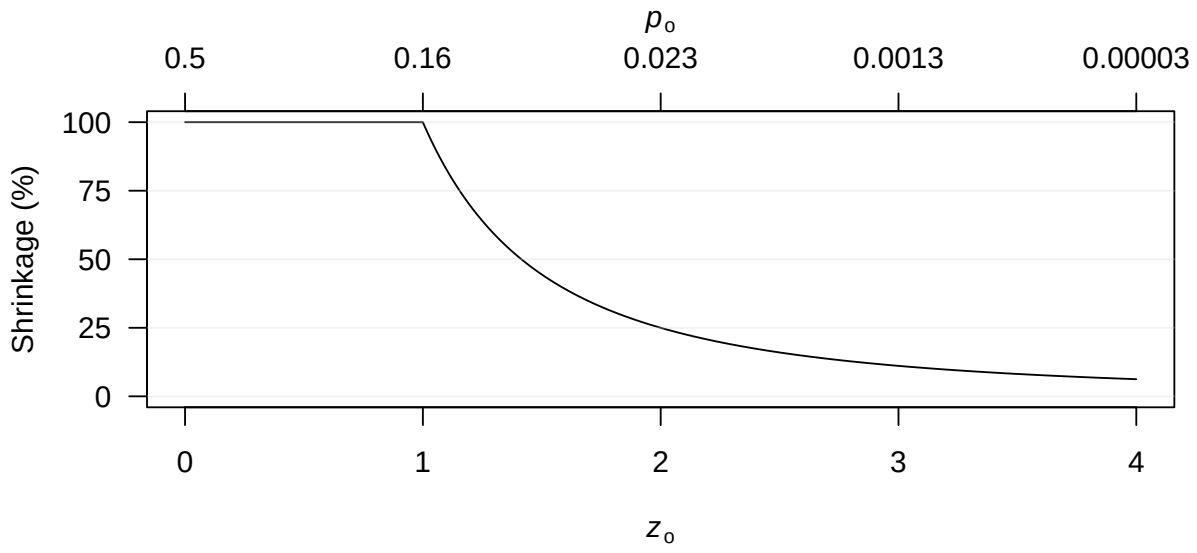


Figure 6: Empirical Bayes shrinkage when there is no between-study heterogeneity.

5 Outlook

Development on `ReplicationSuccess` will continue. We invite anyone with ideas for new functionality, bug-reports, or other contributions to the package to get in touch with us over GitHub (https://github.com/crsuzh/ReplicationSuccess/).

References

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A*, 143:383–430. doi:10.2307/2982063.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (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., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikenstein, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E., and Wu, H. (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.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., van Dongen, N. N. N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., Jaquet, F., Khalifa, K., Kim, H., Kneer, M., Knobe, J., Kurthy, M., Lantian, A., yi Liao, S., Machery, E., Moerenhout, T., Mott, C., Phelan, M., Phillips, J., Rambharose, N., Reuter, K., Romero, F., Sousa, P., Sprenger, J., Thalabard, E., Tobia, K., Viciana, H., Wilkenfeld, D., and Zhou, X. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*. doi:10.1007/s13164-018-0400-9.

Dreber, A., Pfeiffer, T., Almenberg, Isaksson, S., J., Wilson, B., Chen, Y., Nosek, B. A., and Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *PNAS*, 112:15343–15347. doi:10.1073/pnas.1516179112.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32.

Goodman, S. N. (1992). A comment on replication, p-values and evidence. *Statistics in Medicine*, 11(7):875–879. doi:10.1002/sim.4780110705.

Hedges, L. V. and Schauer, J. M. (2019). More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics*, 44(5):543–570. doi:10.3102/1076998619852953.

Held, L. (2019). The assessment of intrinsic credibility and a new argument for $p < 0.005$. *Royal Society Open Science*, 6(3):181534. doi:10.1098/rsos.181534.

Held, L. (2020a). The harmonic mean χ^2-test to substantiate scientific findings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(3):697–708. doi:10.1111/rssc.12410.

Held, L. (2020b). A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(2):431–448. doi:10.1111/rssa.12493.

Held, L., Micheloud, C., and Pawel, S. (2022). The assessment of replication success based on relative effect size. *The Annals of Applied Statistics*, 16:706–720. URL https://doi.org/10.1214/21-AOAS1502.

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., and Mandal, S. (2016). On the reproducibility of psychological science. *Journal of the American Statistical Association*, 112(517):1–10. doi:10.1080/01621459.2016.1240079.

Matthews, R. A. J. (2001). Methods for assessing the credibility of clinical trial outcomes. *Drug Information Journal*, 35:1469–1478. doi:10.1177/009286150103500442.

Micheloud, C., Balabdaoui, F., and Held, L. (2023). Assessing replicability with the sceptical p-value: Type-I error control and sample size planning. *Statistica Neerlandica*. doi:10.1111/stan.12312.

Micheloud, C. and Held, L. (2022). Power calculations for replication studies. *Statistical Science*, 37(3). doi:10.1214/21-STS828.

Morewedge, C. K., Huh, Y. E., and Vosgerau, J. (2010). Thought for food: Imagined consumption reduces actual consumption. *Science*, 330(6010):1530–1533. doi:10.1126/science.1195701.

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

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

Pawel, S. and Held, L. (2020). Probabilistic forecasting of replication studies. *PLOS ONE*, 15(4):e0231416. doi:10.1371/journal.pone.0231416.

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., and et al. (2020). High replicability of newly-discovered social-behavioral findings is achievable. URL psyarxiv.com/n2a9x.