# Predictive Evaluation of Replication Studies

Master Thesis in Biostatistics (STA495)

by

Samuel Pawel
samuel.pawel@uzh.ch
14 – 702 – 823

supervised by

Prof. Dr. Leonhard Held
Department of Biostatistics
University of Zurich


University of Zurich[UZH]

Zurich, August 2019

# Abstract

Throughout the last decade, the so-called replication crisis has stimulated many researchers to conduct large-scale replication projects. With data from four of these projects, we computed probabilistic predictions of the replication outcomes, which we then evaluated regarding discrimination, calibration and sharpness. By using a model of effect sizes which can take into account possible inflation and heterogeneity of effects, it was possible to predict the effect estimate of the replication study with good performance in two of the four data sets. In the other two data sets, predictive performance could still be substantially improved compared to the naive model which does not consider inflation and heterogeneity of effects. The results suggest that many of the estimates from the original studies were too optimistic, possibly caused by publication bias or questionable research practices. Moreover, the results also indicate that the use of statistical significance as the only criterion for replication success may be questionable, since from a predictive viewpoint non-significant replication results are often in agreement with significant results from the original study. Finally, the proposed model could be used to determine the sample size of a new replication study, considering potentially inflated and heterogeneous effect estimates, which seems realistic in view of our results.

# Acknowledgments

Working on this thesis has been an incredibly rewarding experience that helped me to grow both personally and as a statistician. I would like to thank my supervisor Leonhard Held for suggesting the very interesting topic, his support and guidance, as well as the opportunity to present the results at the Bayesian biostatistics conference 2019 in Lyon. I also want to thank all my friends from the Master Program in Biostatistics, in particular Bálint Tamási, Charlotte Micheloud, Eleftheria Michalopoulou, Giuachin Kreiliger, Marielena Syleouni, Peter Meili, Sandar Lim, and Sandra Siegfried. The lectures and studying became even more fun with you. Furthermore, I want to thank all the people involved in the Master Program in Biostatistics, especially Eva and Reinhard Furrer, Leonhard Held, and Torsten Hothorn. I am deeply grateful that you have granted me the opportunity to do this master's and I have enjoyed all of your lectures very much. Also many thanks to my good friends Daniel Fuchs and Mirela Zrnic, you always manage to bring me back to the non-statistical reality. Finally, I would like to thank my parents for their continued support.

Samuel Pawel
August 8, 2019

# Contents

# Chapter 1

# Introduction

Direct replication of past experiments is an essential tool in the modern scientific process for assessing the credibility of scientific discoveries. That is, if a claimed discovery is indeed true, a similar result should be obtained by repeating the original experiment. If the original claim is false, however, one would expect the replication experiments to lead to contradictory results. Moreover, replication is sometimes also a regulatory requirement. For instance, the "two pivotal study paradigm" of the FDA requires statistically significant results from two independent confirmatory trials to grant drug approval (Lee, 2018).

Over the course of the last decade, concerns regarding the replicability of scientific discoveries have increased dramatically, leading many to conclude that science is in a crisis (Ioannidis, 2005; Gelman and Loken, 2014). For this reason, researchers in different fields, *e. g.* psychology or economics, have joined forces to conduct large-scale replication projects. Usually, in such a replication project, representative original studies are carefully selected and then direct replication studies of these original studies are carried out. In a direct replication study, the experimental design is matched as closely as possible to the original study in order to assess the credibility of the original study results. By now, some of these projects have been completed and their data made available to the public, *e. g.* Klein *et al.* (2014); Open Science Collaboration (2015); Ebersole *et al.* (2016); Camerer *et al.* (2016, 2018); Cova *et al.* (2018); Klein *et al.* (2018). The low rate of replication success in some of these projects has received enormous attention in the media and science communities. Moreover, these results lead to an increased awareness of the replication crisis as well as to increased interest in research on the scientific process itself (*meta-science*).

Despite the fact that most researchers agree on the importance of direct replication studies, there is currently no agreement on a universal statistical criterion for replication success. First, statistical significance is commonly used but criticized for many reasons. For example, non-significant replication results are expected if the original finding was a false positive (*e. g.* with 95% probability if the significance level is 5%), on the other hand they are also expected with non-negligible probability if the underlying effect is present (Goodman, 1992; Killeen, 2006; Simonsohn, 2015). Second, the effect estimates of original and replication study are often compared, for instance by examining whether the replication effect estimate is within its 95% prediction interval based on the original effect estimate (Patil *et al.*, 2016) or whether the original effect estimate is within the 95% confidence interval of the replication effect estimate (Open Science Collaboration, 2015). However, for studies which are underpowered (as it is often the case), the confidence and prediction intervals will become very wide. This in turn can lead to the very different effect estimates being compatible, *e. g.* even ones that go strongly in the opposite direction, ultimately providing no information about the effect (Patil *et al.*, 2016). Third, original and replication effect estimates can be combined using meta-analysis methods. By conducting a replication study, however, researchers want to assess the credibility of the original study results in light of the results from the replication study. Combining the effect estimates from both studies and treating them as exchangeable is not a sensible way to answer this question (Held, 2019a).

Fourth, Bayesian hypothesis testing has been proposed to quantify the evidence for the existence of the original effect estimate against the null hypothesis of no effect given the results of the replication study (Verhagen and Wagenmakers, 2014; Ly *et al.*, 2018). These approaches have many attractive properties, *e. g.* one can quantify the evidence also in favor of the null hypothesis. Although Bayesian methods have become increasingly popular in recent years, many applied researchers lack the statistical training to confidently apply them, and therefore still prefer to use frequentist methods for their analyses. Finally, Held (2019a) recently proposed a reverse Bayes approach that tries to address the shortcomings of the above mentioned methods. That is, replication success is quantified by the conflict between the replication effect estimate and a prior predictive distribution which is determined such that after observing the original study, the credible interval of the posterior distribution of the effect size just includes zero. This promising method provides a theoretically sound approach to quantify replication success. However, Held (2019a) also showed that it is a more stringent criterion compared to statistical significance and therefore requires larger samples to achieve replication success. This could make this method unattractive for researchers with little resources for replication studies. The discussion about a statistical criterion for replication success is far from over, it will remain interesting to follow the coming developments.

As already mentioned, to quantify the agreement between original and replication study, Patil *et al.* (2016) introduced the approach of computing a prediction interval of the effect estimate of the replication study based on the effect estimate from the original study and knowledge of the sample size in both studies. Patil *et al.* (2016) used the data set from the replication project psychology (Open Science Collaboration, 2015) to illustrate their method, and the same method was again used in the analyses of the experimental economics replication project (Camerer *et al.*, 2016) and the social sciences replication project (Camerer *et al.*, 2018). In all of these analyses, the coverage of the 95% prediction intervals was examined to assess the predictive performance. Although the assessment of prediction interval coverage may provide some clues about the calibration of the predictions, there exists a whole catalogue of theoretically well-founded methods for the evaluation of probabilistic predictions that are more suitable for this task (for an introduction see *e. g.* Gneiting and Katzfuss, 2014).

From this starting point, this master thesis has several objectives. First, predictions of replication study outcomes based on data from some replication projects, *i. e.* Open Science Collaboration (2015); Camerer *et al.* (2016, 2018); Cova *et al.* (2018), will be computed and systematically evaluated using established methods from the statistical prediction literature. The second goal is then to improve these predictions. Namely, the prediction model used by Patil *et al.* (2016) assumes that the original study correctly identified the underlying effect size, but it is often likely that effect estimates of original studies are inflated, *e. g.* by the influence of publication bias (Dwan *et al.*, 2013; Kicinski *et al.*, 2015) or questionable research practices (Fanelli, 2009; John *et al.*, 2012). Another concern with the model by Patil *et al.* (2016) is that it also assumes the effect estimates from both studies to be realizations of the same underlying effect size. However, it may also be the case that there is between study heterogeneity of the underlying effects (Gilbert *et al.*, 2016; McShane *et al.*, 2019). This can be caused, for example, by different populations of study participants or different laboratory equipment being used in original and replication study. For this reason, a model of effect sizes for the setting of replication studies will be developed which can take into account possible inflation as well as possible between study heterogeneity of effect estimates.

The structure of this thesis is as follows. First, in the methods chapter various methods for obtaining and evaluating probabilistic predictions in the setting of replication studies are discussed, additionally the used data sets are described. Second, in the results chapter, descriptive results about the data sets as well as results from the evaluation of the predictions are summarized. Finally, the thesis ends with a discussion of these results and closing conclusions, extensions, and limitations.

# Chapter 2

# Methods

## 2.1 General framework

In this section notation is introduced and some general results are established which will be used subsequently. Unless otherwise stated, they are taken from Held and Sabanés Bové (2014).

### 2.1.1 Normal distribution

Let $x$ be a realization of a random variable $X$ that follows a normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_{>0}$, which is commonly abbreviated by $X \sim \mathrm{N}(\mu, \sigma^2)$. The probability density function of $x$ is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}.$$

$X$ is said to be standard normal if $\mu = 0$ and $\sigma^2 = 1$. Throughout the whole thesis, it is assumed that after suitable transformations, an effect size $\theta$ can be modelled by a normally distributed random variable with known variance. This framework supports a wide range of commonly used effect sizes, for example, mean differences, odds ratios, correlations, or hazard ratios.

### 2.1.2 Maximum likelihood estimation

The likelihood function, $L(\theta)$, is the probability mass or probability density function of the data $x$ as a function of the unknown parameter $\theta$. The maximum likelihood estimate, $\hat{\theta}_{\mathrm{ML}}$, is then obtained by maximizing the (log-) likelihood function

$$\begin{aligned}\hat{\theta}_{\mathrm{ML}} &= \arg\max_{\theta} L(\theta) = \arg\max_{\theta} f(x \mid \theta) \\ &= \arg\max_{\theta} \log f(x \mid \theta).\end{aligned}$$

If $x$ is a realization from $X \sim \mathrm{N}(\mu, \sigma^2)$ with $\sigma^2$ known, the log-likelihood function, $l(\mu)$, is given by

$$l(\mu) = \log f(x \mid \mu) = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}.$$

Differentiating with respect to $\mu$, setting the equation to zero and solving for $\mu$ yields $\hat{\mu}_{\mathrm{ML}} = x$.

### 2.1.3 Bayes' theorem and posterior distribution

In the Bayesian framework, the unknown parameter $\theta$ is not fixed but itself a random variable following a *prior distribution* with probability density or mass function $f(\theta)$. After having

observed realization $x$ of a random variable $X$ with density $f(x \mid \theta)$, the density $f(\theta \mid x)$ of the *posterior distribution* can be computed by using Bayes' theorem

$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{f(x)} = \frac{f(x \mid \theta) f(\theta)}{\int f(x \mid \theta) f(\theta) d\theta},$$

where the denominator, $f(x) = \int f(x \mid \theta) f(\theta) d\theta$, is known as the *marginal likelihood*.

For the model, where $x$ is a realization from $X \sim \mathrm{N}(\mu, \sigma^2)$ with $\sigma^2$ known and prior distribution $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$, the posterior density of $\mu$ is

$$\begin{aligned} f(\mu \mid x) &= \frac{f(x \mid \mu) f(\mu)}{\int f(x \mid \mu) f(\mu) d\mu} \\ &\propto \exp\left[ -\frac{1}{2} \left\{ \frac{(x-\mu)^2}{\sigma^2} + \frac{(\mu - \mu_0)^2}{\sigma_0^2} \right\} \right] \\ &\propto \exp\left[ -\frac{1}{2} \left\{ \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} \right\} \left( \mu - \left\{ \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} \right\}^{-1} \cdot \left\{ \frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right\} \right)^2 \right], \end{aligned}$$

which, after some algebraic rearrangements, can be identified as the kernel of another normal distribution, namely the posterior distribution of $\mu$ is

$$\mu \mid x \sim \mathrm{N}\left( \left\{ \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} \right\}^{-1} \cdot \left( \frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right), \left\{ \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} \right\}^{-1} \right). \tag{2.1}$$

Similarly, the marginal likelihood is

$$\begin{aligned} f(x) &= \int f(x \mid \mu) f(\mu) d\mu \\ &\propto \int \exp\left[ -\frac{1}{2} \left\{ \frac{(x-\mu)^2}{\sigma^2} + \frac{(\mu - \mu_0)^2}{\sigma_0^2} \right\} \right] d\mu \\ &\propto \exp\left[ -\frac{1}{2} \frac{(x - \mu_0)^2}{\sigma^2 + \sigma_0^2} \right]. \end{aligned}$$

After some algebraic rearrangements and by using that for $a > 0$, $\int \exp(-ax^2) dx = \sqrt{\pi / a}$, the expression can be identified as the kernel of a normal density, in particular

$$x \sim \mathrm{N}\left( \mu_0, \sigma^2 + \sigma_0^2 \right). \tag{2.2}$$

### 2.1.4 Posterior predictive distribution

If a realization $x$ from the random variable $X$ following a distribution with probability density $f(x \mid \theta)$ is observed and the goal is to predict a new observation $Y$ also with density $f(x \mid \theta)$, the posterior predictive density $f(y \mid x)$ can be derived to be

$$\begin{aligned} f(y \mid x) &= \int f(y, \theta \mid x) d\theta = \int f(y \mid \theta, x) f(\theta \mid x) d\theta \\ &= \int f(y \mid \theta) f(\theta \mid x) d\theta. \end{aligned}$$

Note that this is exactly the same expression as the marginal likelihood, just with the density of the prior distribution $f(\theta)$ replaced by the density of the posterior distribution $f(\theta \mid x)$.

By using (2.1) and (2.2), the posterior predictive distribution for the normal model with known variance becomes

$$y \mid x \sim \mathrm{N}\left( \left\{ \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} \right\}^{-1} \cdot \left( \frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right), \left\{ \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} \right\}^{-1} + \sigma^2 \right). \tag{2.3}$$

### 2.1.5 Effect size scale

In the analysis of data from replication projects it has become common practice to transform effect sizes to the correlation coefficient scale $\rho$ (Open Science Collaboration, 2015; Camerer *et al.*, 2016, 2018; Cova *et al.*, 2018). For an introduction to the conversion between effect size scales, see section 12.5 in Cooper *et al.* (2009). An advantage of correlation coefficients is that they are bounded to the interval between minus one and one and are thus easy to compare and interpret. Moreover, by applying the variance stabilizing transformation, also known as Fisher $z$-transformation, $z(\rho) = \tanh^{-1}(\rho)$, the transformed correlation coefficients become asymptotically normally distributed with their variance only being a function of the study sample size $n$, *i. e.* $\text{Var}(z(\hat{\rho})) = 1/(n-3)$ (Fisher, 1921). The Fisher $z$-transformation is shown in Figure 2.1. Throughout this thesis modelling and prediction will be carried out on the Fisher $z$ scale, but the results will often be backtransformed to the correlation scale by applying the inverse Fisher $z$-transformation, $\rho = \tanh(z)$, for better comparability and interpretability.

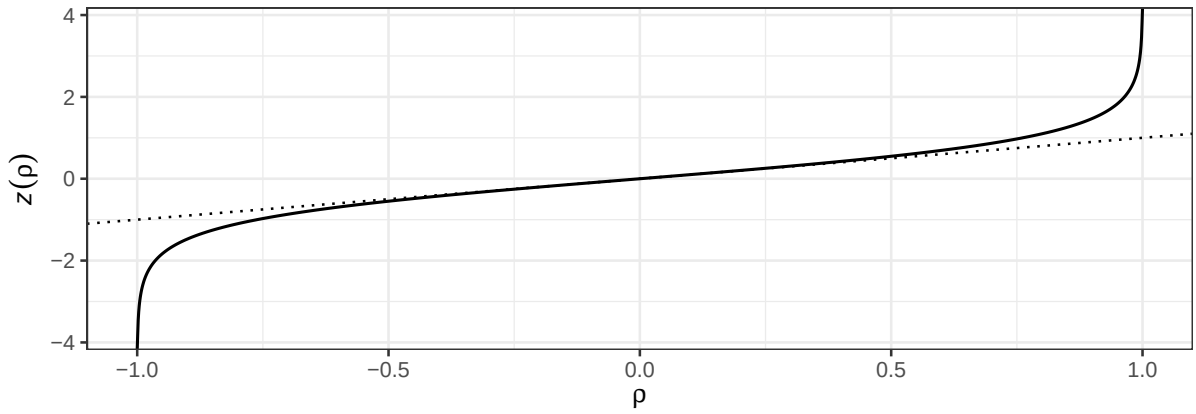


**Figure 2.1:** Fisher $z$-transformation.

Open Science Collaboration (2015) used the approach of computing *correlation per degree of freedom* based on the test statistics of the original effect estimates (see page 74 in the supplementary material of Open Science Collaboration, 2015). This is possible for $z$, $\chi^2$, $t$, and $F$ test statistics and can be done using the following formulas

$$\begin{aligned}
\rho(t) &= \frac{\sqrt{t^2/\text{df}_2}}{(t^2/\text{df}_2)+1} \\
\rho(F) &= \frac{\sqrt{F(\text{df}_1/\text{df}_2)}}{\{F(\text{df}_1/\text{df}_2)+1\}\sqrt{1/\text{df}_1}} \\
\rho(\chi^2) &= \sqrt{\chi^2/n} \\
\rho(z) &= \tanh\left(z\sqrt{\frac{1}{n-3}}\right).
\end{aligned}$$

The approach has become the standard for further replication projects (Camerer *et al.*, 2016, 2018; Cova *et al.*, 2018).

### 2.1.6 Notation

Throughout this thesis $\hat{\theta}_o$ and $\hat{\theta}_r$ denote the effect estimates after suitable transformation with their subscript indicating whether they come from the original or the replication study. The corresponding standard errors are denoted by $\sigma_o$ and $\sigma_r$ and assumed to be known. Similarly, define the variance ratio as $c = \sigma_o^2/\sigma_r^2$ and also define the test statistics $t_o$ and $t_r$ obtained by

dividing the effect estimate by its standard error, *e. g.* $t_o = \hat{\theta}_o/\sigma_o$. Also let $\rho$ denote the Pearson correlation coefficient applied to a population, and let a sample-based estimate of it be denoted by $\hat{\rho} = r$. Finally, let $\Phi(x)$ and $\varphi(x)$ be the cumulative distribution and probability density function of the standard normal distribution evaluated at $x$ and let $z_\alpha$ denote the $1-\alpha$ quantile thereof.

## 2.2 Prediction methods

In this section, methods are discussed which allow to compute probabilistic predictions of the replication study outcome based on the original study outcome. The methods were implemented in R and the corresponding code can be found in Appendix A.1

### 2.2.1 Plug-in predictive distribution

The plug-in predictive density is obtained by replacing the unknown parameter in the density function underlying the data with its estimate, *e. g.* the maximum likelihood estimate $\hat{\theta}_{\text{ML}}$ (Held and Sabanés Bové, 2014).

If the effect estimates are assumed to be normally distributed, *e. g.* $\hat{\theta}_r \sim \text{N}(\theta, \sigma_r^2)$, the plug-in predictive distribution of the replication study effect estimate becomes

$$\hat{\theta}_r \,|\, \hat{\theta}_o \sim \text{N}(\hat{\theta}_o, \sigma_r^2).$$

Given this predictive model, the distribution of the test statistic of the replication study is $t_r \,|\, \hat{\theta}_o \sim \text{N}(\hat{\theta}_o/\sigma_r, 1)$. Note that $\hat{\theta}_o/\sigma_r = \hat{\theta}_o/\sigma_o \cdot \sigma_o/\sigma_r = t_o\sqrt{c}$, and hence, the probability of a statistically significant replication outcome at the $\alpha$ level and with the same sign as the original effect estimate is

$$\text{Pr}(t_r > z_{\alpha/2} \,|\, \hat{\theta}_o) = 1 - \Phi\left(z_{\alpha/2} - \hat{\theta}_o/\sigma_r\right) = \Phi\left(t_o\sqrt{c} - z_{\alpha/2}\right),$$

which in the context of sample size planning is known as the classical power (Spiegelhalter *et al.*, 2004). However, in this model the uncertainty with respect to estimating $\theta$ is ignored, resulting in inferior predictive performance compared to the methods discussed further below. The plug-in method is only mentioned, to serve as a benchmark for the other methods and because it is commonly used for sample size calculations (despite taking the uncertainty not properly into account).

### 2.2.2 Posterior predictive distribution

**Flat prior** If the prior distribution of $\theta$ is chosen to be flat, *i. e.* $\theta \sim \text{N}(0, \infty)$, with (2.1) the posterior distribution of the effect size $\theta$ after observing the original study effect estimate $\hat{\theta}_o$ becomes

$$\theta \,|\, \hat{\theta}_o \sim \text{N}(\hat{\theta}_o, \sigma_o^2).$$

By using (2.3), the posterior predictive distribution of $\hat{\theta}_r$ can then be identified to be

$$\hat{\theta}_r \,|\, \hat{\theta}_o \sim \text{N}(\hat{\theta}_o, \sigma_o^2 + \sigma_r^2). \tag{2.4}$$

Hence, by using this predictive model, one assumes the original study to correctly identify the true effect size. In contrast to the plug-in predictive distribution, the uncertainty of $\hat{\theta}_o$ is also taken into account.

Under this predictive distribution, the distribution of the test statistic of the replication study is $t_r \,|\, \hat{\theta}_o \sim \mathrm{N}(t_o\sqrt{c}, c+1)$ and hence the probability of a statistically significant replication outcome at the $\alpha$ level and with the same sign as the original effect estimate is

$$\mathsf{Pr}(t_r > z_{\alpha/2} \,|\, \hat{\theta}_o) = \Phi\left(\frac{t_o\sqrt{c} - z_{\alpha/2}}{\sqrt{c+1}}\right).$$

In the context of sample size planning, this quantity is sometimes called hybrid power (Spiegelhalter *et al.*, 2004) or predictive power (Rufibach *et al.*, 2016) since a Bayesian prediction for the outcome of a frequentist analysis is performed. For brevity and in analogy to the existing naming convention, this method will be referred to as the *predictive* method, although the other methods also provide a predictive distribution.

**Sceptical prior** It is also possible to choose a different prior distribution for the effect size $\theta$, reflecting a more sceptical belief about $\theta$. That is, if Zellner's $g$-prior (Zellner, 1986) is chosen, *i. e.* $\theta \sim \mathrm{N}(0, g \cdot \sigma_o^2)$ with $g \geq 0$, when marginalizing over $\theta$ by using (2.2), the marginal distribution of $\hat{\theta}_o$ becomes

$$\hat{\theta}_o \sim \mathrm{N}(0, (1+g) \cdot \sigma_o^2).$$

A well-founded approach to specify the parameter $g$ when no prior knowledge is available, is to choose it such that the marginal likelihood is maximized (known as empirical Bayes estimation). The marginal log-likelihood is given by

$$\begin{aligned} l(g) &= \log\left(\left\{2\pi(1+g)\sigma_o^2\right\}^{-1/2} \exp\left\{-\frac{1}{2}\frac{\hat{\theta}_o^2}{(1+g)\sigma_o^2}\right\}\right) \\ &= -\frac{1}{2}\log(2\pi\sigma_o^2) - \frac{1}{2}\log(1+g) - \frac{1}{2}\frac{t_o^2}{(1+g)}. \end{aligned}$$

Differentiating $l(g)$ with respect to $g$ leads to

$$\begin{aligned} \frac{dl(g)}{dg} &= -\frac{1}{2}\left\{\frac{1}{1+g} - \frac{t_o^2}{(1+g)^2}\right\} \\ &= \frac{1}{2}\frac{1}{1+g}\left\{-1 + \frac{t_o^2}{(1+g)}\right\}. \end{aligned}$$

By equating the expression to zero and solving for $g$, the empirical Bayes estimate

$$\hat{g} = \max\left\{t_o^2 - 1, 0\right\}$$

is obtained.

Fixing $g$ to $\hat{g}$ and using (2.1), the posterior distribution of the effect size $\theta$ after observing the original study effect estimate $\hat{\theta}_o$ can be identified as

$$\theta \,|\, \hat{\theta}_o, \hat{g} \sim \mathrm{N}(s \cdot \hat{\theta}_o, s \cdot \sigma_o^2),$$

with shrinkage factor

$$s = \frac{\hat{g}}{\hat{g}+1} = \max\left\{1 - \frac{1}{t_o^2}, 0\right\}.$$

Figure 2.2 shows the shrinkage factor $s$ as function of the test statistic respectively of the two-sided $p$-value of the original study. Interestingly, this is a special case of the shrinkage factor of the Stein-type predictor derived by Copas (1983) in a regression setting and which was shown

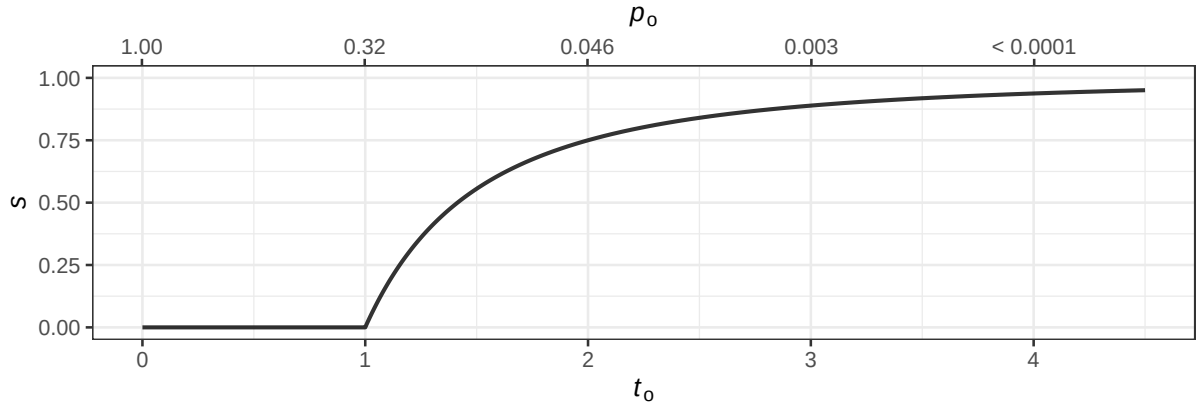


**Figure 2.2:** Shrinkage factor $s$ as function of the test statistic $t_o$ respectively of the two-sided $p$-value $p_o$ of the original study.

to give uniformly lower prediction mean squared error compared to least squares. The same shrinkage factor was again discussed by Copas (1997) in the context of overcoming the effect of regression to the mean in prediction problems. Shrinkage proved to be particularly useful when the number of covariates in the regression model is large and/or the size of the sample used to fit the model is small, as it is the case for the current replication setting where only one original study is observed.

Using (2.3), the posterior predictive distribution of $\hat{\theta}_r$ is derived to be

$$\hat{\theta}_r \,|\, \hat{\theta}_o \sim \mathrm{N}(s \cdot \hat{\theta}_o, s \cdot \sigma_o^2 + \sigma_r^2). \tag{2.5}$$

Thus, when using this predictive model, the original study effect estimate and its variance are shrunken towards zero depending on the amount of evidence (*evidence based shrinkage*). If there was substantial evidence for an effect, *i. e.* $t_o$ was very large or $p_o$ was very small, the shrinkage factor $s \approx 1$ and hence the predictive distribution is virtually identical to the predictive distribution when using a flat prior for $\theta$. However, if the evidence in the original study was only suggestive, *i. e.* $t_o \approx 2$ or $p_o \approx 0.05$, the effect estimate from the original study as well as the corresponding variance term are shrunken towards zero, leading to a less optimistic prediction compared to when using a flat prior for $\theta$.

Based on this predictive distribution, the test statistic of the replication study is distributed as $t_r \,|\, \hat{\theta}_o \sim \mathrm{N}(s \cdot t_o\sqrt{c}, s \cdot c + 1)$, which leads to the probability of a statistically significant replication outcome at the $\alpha$ level and with the same sign as the original effect estimate being

$$\mathrm{Pr}(t_r > z_{\alpha/2} \,|\, \hat{\theta}_o) = \begin{cases} \Phi\left(\dfrac{t_o\sqrt{c} - \sqrt{c}/t_o - z_{\alpha/2}}{\sqrt{c + 1 - c/t_o^2}}\right) & \text{if } t_o^2 > 1 \\ \Phi\left(-z_{\alpha/2}\right) & \text{if } t_o^2 \leq 1. \end{cases}$$

If there is hardly any evidence for an effect in the original study, *i. e.* $|t_o| \leq 1$, the predicted probability of a significant replication outcome is just the type I error $\alpha$. On the other hand, in the limiting case when $t_o \to \infty$, this probability is the same as when using a flat prior for the effect size $\theta$. In the remaining part of the thesis, this prediction method will be referred to as the *shrinkage* method for brevity reasons.

### 2.2.3 Taking into account between study heterogeneity

Even in direct replication studies, where the conditions of original and replication study are as closely matched as possible, it is very likely that there is natural between study heterogeneity of

the underlying effects. For example, the original and replication study might have been conducted using slightly different populations of participants or different laboratory equipment. This is a common objection against the validity of results from replication studies (*e. g.* in Gilbert *et al.*, 2016).

One way of incorporating between study heterogeneity into the current objective is by assuming a hierarchical model of the effect size parameters, *i. e.*

$$\begin{aligned} \theta &\sim \mathrm{N}(\mu_\theta, \sigma_\theta^2) \\ \theta_k \,|\, \theta &\sim \mathrm{N}(\theta, \tau^2) \\ \hat{\theta}_k \,|\, \theta_k &\sim \mathrm{N}(\theta_k, \sigma_k^2), \end{aligned}$$

where $k \in \{o, r\}$ and $\tau^2$ is the heterogeneity variance (see Figure 2.3 for a graphical illustration).

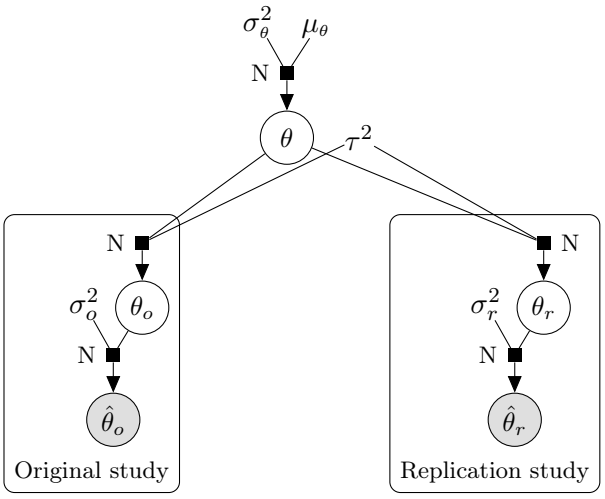


**Figure 2.3:** Hierarchical model of effect size parameters.

Marginalizing over $\theta_k$ with (2.2) leads to the marginal distribution of $\hat{\theta}_k$ being

$$\hat{\theta}_k \,|\, \theta \sim \mathrm{N}(\theta, \sigma_k^2 + \tau^2),$$

which can be used as in the previous derivations to obtain the posterior distribution of $\theta$ and the posterior predictive distribution of $\hat{\theta}_r$ given the observed effect estimate of the original study $\hat{\theta}_o$.

**Flat prior** If a flat prior for $\theta$ is chosen, *i. e.* $\theta \sim \mathrm{N}(0, \infty)$, using (2.1), the posterior distribution of $\theta$ after observing the original study effect estimate $\hat{\theta}_o$ becomes

$$\theta \,|\, \hat{\theta}_o \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2 + \tau^2).$$

As before, with (2.3) the predictive distribution of $\hat{\theta}_r$ given $\hat{\theta}_o$ can be derived to be

$$\hat{\theta}_r \,|\, \hat{\theta}_o \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2 + \sigma_r^2 + 2\tau^2). \tag{2.6}$$

Hence, by using this predictive model, one expects the original study to identify the effect size correctly. Furthermore, the uncertainty coming from original and replication study, as well as the uncertainty from the between study heterogeneity is taken into account.

Given this predictive distribution, the test statistic of the replication study given the results of the original study is distributed as $t_r \,|\, \hat{\theta}_o \sim \mathrm{N}(t_o\sqrt{c}, c + 1 + 2\tau^2/\sigma_r^2)$. Now define the *relative between study heterogeneity* as $d = \tau^2/\sigma_o^2$, the ratio of the heterogeneity variance to the squared standard error of the original study effect estimate, and note that $\tau^2/\sigma_r^2 = \tau^2/\sigma_o^2 \cdot \sigma_o^2/\sigma_r^2 = d \cdot c$.

The probability of a statistically significant replication outcome at the $\alpha$ level and with the same sign as the original effect estimate is then

$$\mathsf{Pr}(t_r > z_{\alpha/2} \,|\, \hat{\theta}_o) = \Phi\left(\frac{t_o\sqrt{c} - z_{\alpha/2}}{\sqrt{c(1+2d)+1}}\right).$$

In the case of no heterogeneity, *i. e.* $d = 0$, this probability reduces to the probability under the standard predictive method.

**Sceptical prior** When choosing again the prior $\theta \sim \mathrm{N}(0, g \cdot \sigma^2)$, marginalizing over $\theta$ by using (2.2) leads to the marginal distribution of $\hat{\theta}_o$ being

$$\hat{\theta}_o \sim \mathrm{N}(0, (1+g) \cdot \sigma_o^2 + \tau^2).$$

Differentiating the marginal log-likelihood with respect to $g$ results in

$$\begin{aligned}
\frac{dl(g)}{dg} &= \frac{d}{dg}\left\{-\frac{1}{2}\log\left\{(1+g)\sigma_o^2+\tau^2\right\} - \frac{1}{2}\frac{\hat{\theta}_o^2}{(1+g)\sigma_o^2+\tau^2}\right\} \\
&= -\frac{1}{2}\left\{\frac{\sigma_o^2}{(1+g)\sigma_o^2+\tau^2} - \frac{\hat{\theta}_o^2\sigma_o^2}{\{(1+g)\sigma_o^2+\tau^2\}^2}\right\} \\
&= -\frac{1}{2}\frac{\sigma_o^2}{(1+g)\sigma_o^2+\tau^2}\left\{1 - \frac{\hat{\theta}_o^2}{(1+g)\sigma_o^2+\tau^2}\right\}.
\end{aligned}$$

Equating this expression to zero and solving for $g$ leads to the empirical Bayes estimate of $g$ being

$$\hat{g} = \max\left\{\frac{\hat{\theta}_o^2 - \tau^2}{\sigma_o^2} - 1, 0\right\} = \max\left\{t_o^2 - d - 1, 0\right\}.$$

Using (2.1) to obtain $\theta \,|\, \hat{\theta}_o, \hat{g} \sim \mathrm{N}(\tilde{\mu}, \tilde{\sigma}^2)$ the posterior distribution of $\theta$ after observing $\hat{\theta}_o$ and setting $g$ to $\hat{g}$, the posterior mean $\tilde{\mu}$ can be identified to be

$$\begin{aligned}
\tilde{\mu} &= \left(\frac{1}{\sigma_o^2+\tau^2} + \frac{1}{\hat{g}\sigma_o^2}\right)^{-1} \cdot \left(\frac{\hat{\theta}_o}{\sigma_o^2+\tau^2} + \frac{0}{\hat{g}\sigma_o^2}\right) \\
&= \frac{\hat{g}\sigma_o^2(\sigma_o^2+\tau^2)}{(\hat{g}+1)\sigma_o^2+\tau^2} \cdot \frac{\hat{\theta}_o}{\sigma_o^2+\tau^2} = \frac{\hat{g}}{\hat{g}+1+d} \cdot \hat{\theta}_o = \tilde{s} \cdot \hat{\theta}_o
\end{aligned}$$

and similarly the posterior variance $\tilde{\sigma}^2$ becomes

$$\tilde{\sigma}^2 = \left(\frac{1}{\sigma_o^2+\tau^2} + \frac{1}{\hat{g}\sigma_o^2}\right)^{-1} = \frac{\hat{g}}{\hat{g}+1+d} \cdot (\sigma_o^2+\tau^2) = \tilde{s} \cdot (\sigma_o^2+\tau^2),$$

where $\tilde{s}$ is a shrinkage factor

$$\tilde{s} = \frac{\hat{g}}{\hat{g}+1+d} = \max\left\{\frac{t_o^2 - d - 1}{t_o^2}, 0\right\} = \max\left\{1 - \frac{1+d}{t_o^2}, 0\right\}.$$

Figure 2.4 illustrates the shrinkage factor $\tilde{s}$ as a function of $t_o$ and $d$. As can be seen, if there is no between study heterogeneity, *i. e.* $d = 0$, the shrinkage factor $\tilde{s}$ reduces to the previously derived shrinkage factor $s$. However, if there is between study heterogeneity, *i. e.* $d > 0$, shrinkage towards zero is not only driven by the evidence in the original study (summarized by $t_o$), but

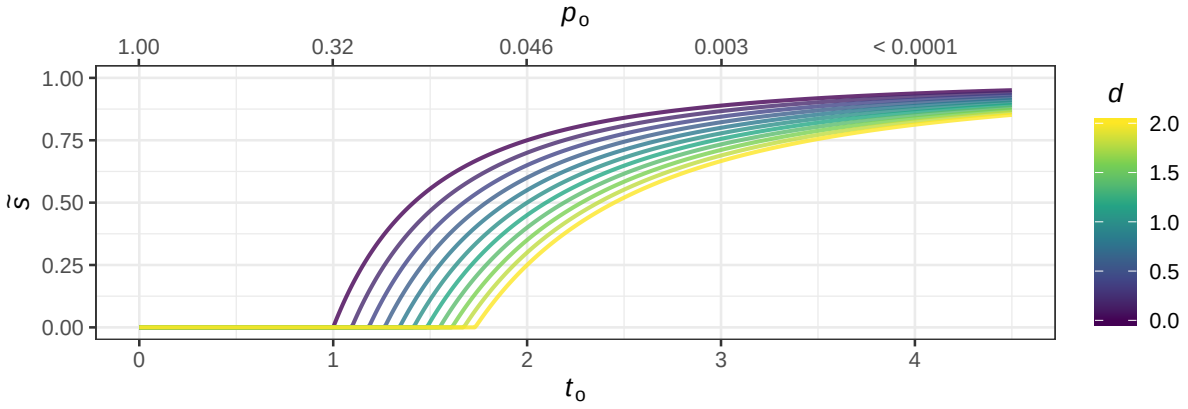


**Figure 2.4:** Shrinkage factor $\tilde{s}$ as function of the test statistic $t_o$ respectively the two-sided $p$-value $p_o$ of the original study and the relative between study heterogeneity $d = \tau^2/\sigma_o^2$.

also by the ratio $d/t_o^2$. If the test statistic is large in comparison to the relative between study heterogeneity, *i. e.* $t_o^2 \gg d$, the contribution of the heterogeneity towards the shrinkage to zero will only be very small. On the other hand, if the size of the test statistic is not substantially larger than the relative between study heterogeneity, shrinkage will also be influenced by $d$.

Remarkably, when choosing instead the prior $\theta \sim \mathrm{N}(0, g \cdot (\sigma_o^2 + \tau^2))$, which corresponds to Zellner's $g$-prior for the marginal likelihood $\hat{\theta}_o \,|\, \theta \sim \mathrm{N}(\theta, \sigma_o^2 + \tau^2)$, the same posterior distribution is obtained when estimating $g$ by empirical Bayes. That is, using the results for Zellner's $g$-prior from the non-heterogeneity setting and just replacing $\sigma_o^2$ by $\sigma_o^2 + \tau^2$, leads to the empirical Bayes estimate of $g$ being

$$\hat{g} = \max\left\{\frac{\hat{\theta}_o^2}{\sigma_o^2 + \tau^2} - 1, 0\right\},$$

and thus the posterior distribution of $\theta$ after observing $\hat{\theta}_o$ turns out to be

$$\theta \,|\, \hat{\theta}_o \sim \mathrm{N}(\tilde{s} \cdot \hat{\theta}_o, \tilde{s} \cdot (\sigma_o^2 + \tau^2)),$$

with shrinkage factor

$$\tilde{s} = \frac{\hat{g}}{1 + \hat{g}} = \max\left\{\frac{\hat{\theta}_o^2/(\sigma_o^2 + \tau^2) - 1}{\hat{\theta}_o^2/(\sigma_o^2 + \tau^2)}, 0\right\} = \max\left\{1 - \frac{1 + d}{t_o^2}, 0\right\},$$

which is the same shrinkage factor $\tilde{s}$ as derived for the prior $\theta \sim \mathrm{N}(0, g \cdot \sigma_o^2)$. Hence, if estimating the $g$ parameter by empirical Bayes, it does not matter which of the two priors is chosen, the same posterior distribution is obtained.

Using (2.2), the posterior predictive distribution of $\hat{\theta}_r$ under this predictive model can be derived to be

$$\hat{\theta}_r \,|\, \hat{\theta}_o \sim \mathrm{N}(\tilde{s} \cdot \hat{\theta}_o, \tilde{s} \cdot (\sigma_o^2 + \tau^2) + \sigma_r^2 + \tau^2). \tag{2.7}$$

Based on this predictive distribution, the test statistic of the replication study is distributed as

$$t_r \,|\, \hat{\theta}_o \sim \mathrm{N}(\tilde{s} \cdot t_o\sqrt{c}, \tilde{s} \cdot (c + dc) + 1 + dc),$$

which leads to the probability of a statistically significant replication outcome at the $\alpha$ level and with the same sign as the original effect estimate being

$$\Pr(t_r > z_{\alpha/2} \mid \hat{\theta}_o) = \begin{cases} \Phi\left(\dfrac{\sqrt{c}(t_o - (1+d)/t_o) - z_{\alpha/2}}{\sqrt{c(1+2d) + 1 - c(1+d)^2/t_o^2}}\right) & \text{if } \dfrac{1+d}{t_o^2} < 1 \\ \Phi\left(\dfrac{-z_{\alpha/2}}{\sqrt{1+dc}}\right) & \text{if } \dfrac{1+d}{t_o^2} \geq 1. \end{cases}$$

Similarly as in the non-heterogeneity case, for $t_o \to \infty$, this probability approaches the probability under the flat prior. Also if there is no heterogeneity, *i. e.* $d = 0$, the probabilities reduce to the probabilities under the standard shrinkage method.

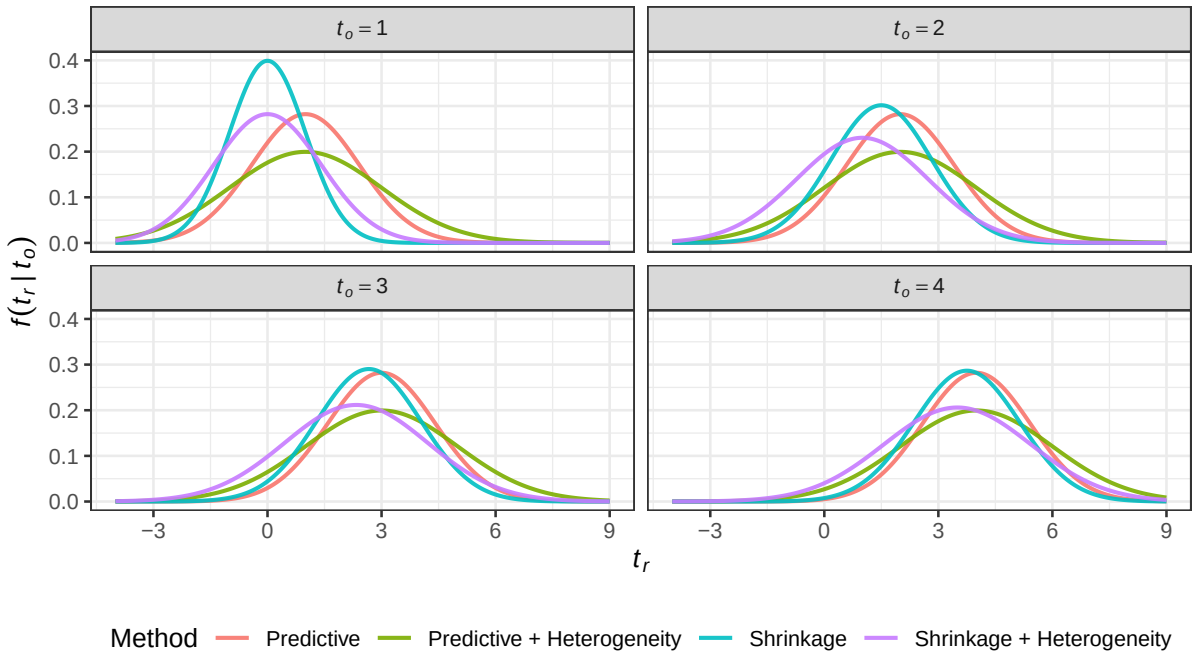


**Figure 2.5:** Comparison of the predictive densities of the discussed prediction methods. In all examples $c = 1$, in the case of heterogeneity $d = 1$.

Figure 2.5 illustrates the predictive densities of the discussed prediction models. As can be seen, when taking into account between study heterogeneity, the predictive densities become wider compared to when not taking into account between study heterogeneity, reflecting the additional uncertainty about the effect size. Furthermore, if there was no convincing evidence for an effect in the original study, *i. e.* $t_o$ was small, the shrinkage predictive densities are substantially shrunken towards zero. For increasing $t_o$, on the other hand, the shrinkage predictive densities approach the predictive densities which arise when choosing a flat prior for the effect size.

Figure 2.6 shows the probability of obtaining a significant effect estimate in the replication study going in the same direction as the effect estimate of the original study as a function of $t_o$ and for different values of $c$. Focusing on $c = 1$, if the original study showed a $p$-value of 0.05, which corresponds to $t_o \approx 1.96$, the probability of repeating a statistical significant result, when assuming the original study correctly identified the effect size, is just about 0.5. This counterintuitive result was already noted by Goodman (1992). For the predictions which are subject to shrinkage, this quantity is even lower. Moreover, if the effect size of the replication is estimated less precisely than in the original study (*i. e.* $c < 1$), the probability of significance in the replication study becomes lower compared to when using the same precision. On the other hand, if the precision of estimating the effect is increased, (*i. e.* $c > 1$), the probability of significance in the replication also increases. Furthermore, for small $t_o$ the probabilities are higher for methods taking into account heterogeneity compared to their counterparts which do

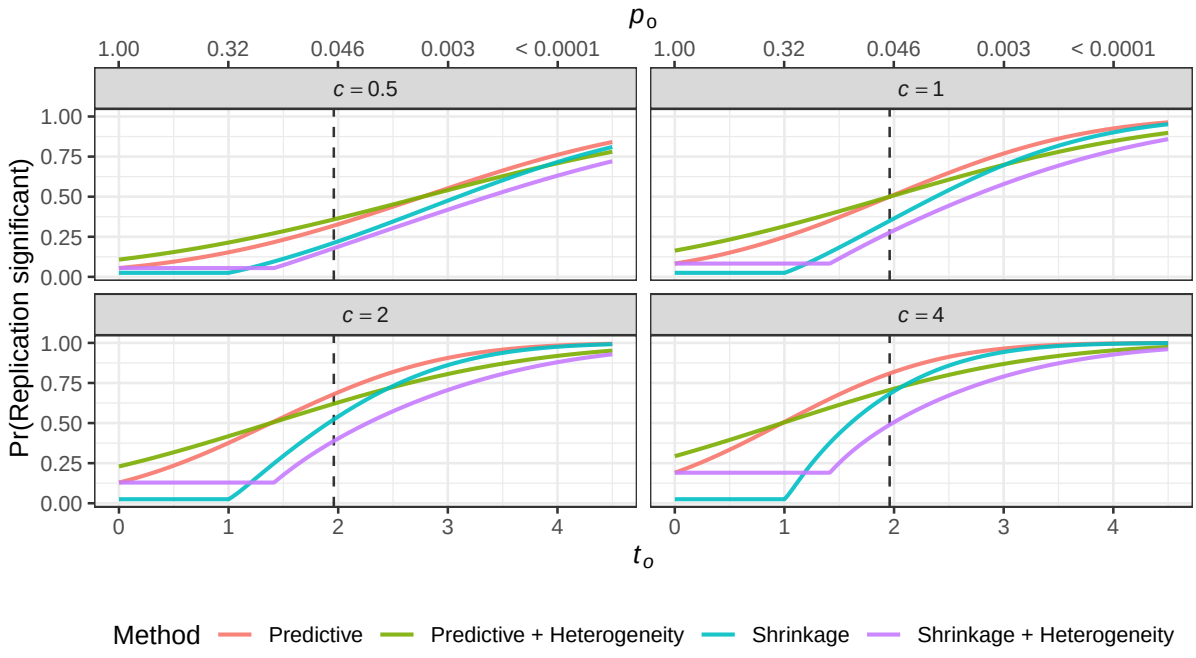


**Figure 2.6:** Probability of a significant replication outcome at $\alpha = 0.05$ as a function of the test statistic $t_o$ or $p$-value $p_o$ of the original study and variance ratio $c$. The dashed line indicates $z_{0.025}$. In the case of heterogeneity, $d = 1$.

not take into account heterogeneity, while the reverse is true for large $t_o$. The $t_o$ at which the change happens depends on $c$ and also differs between the shrinkage and the predictive method.

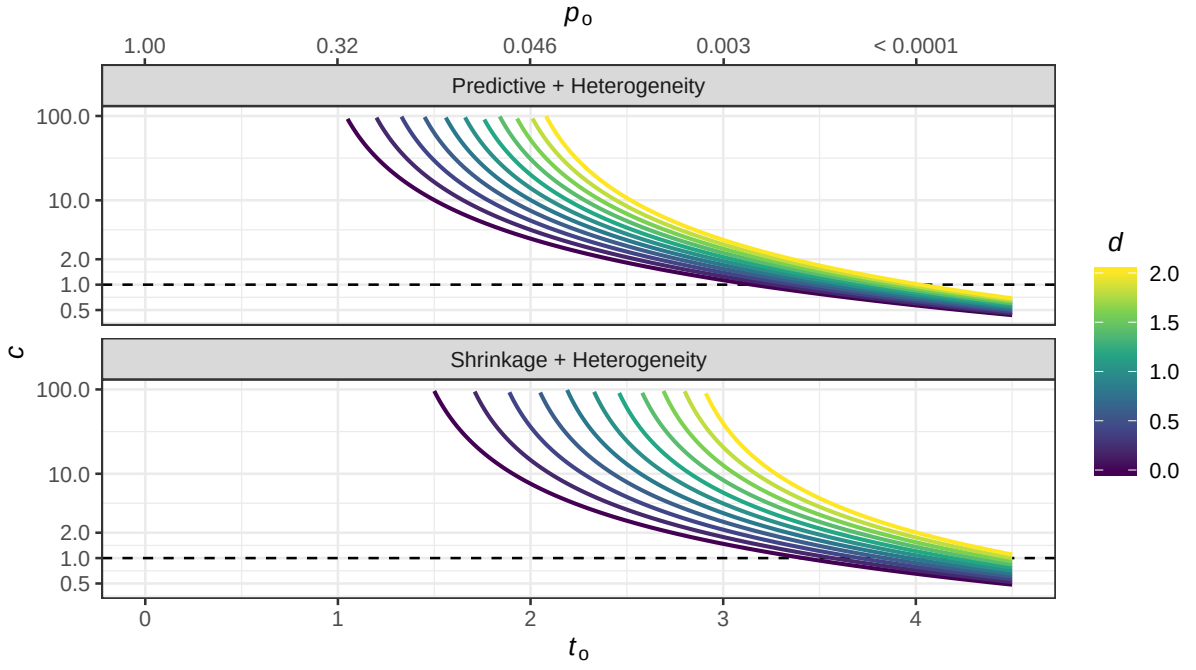


**Figure 2.7:** Required relative sample size $c = n_r/n_o$ to achieve a power of 80% as a function of the test statistic $t_o$ respectively the two-sided $p$-value $p_o$ of the original study and the relative between study heterogeneity $d$.

Assuming that the standard errors of the effect estimates only depend on some unit variance

$\kappa^2$ and the sample size of the study, *i. e.* $\sigma_o^2 = \kappa^2/n_o$ and $\sigma_r^2 = \kappa^2/n_r$, the required relative sample size $c = n_r/n_o$ to achieve a statistically significant result in the replication study with a certain power can be computed using root-finding algorithms. In Figure 2.7, the required $c$ to achieve 80% power under the different models is shown as a function of $t_o$ and the relative between study heterogeneity $d$. As can be seen, the required relative sample size $c$ decreases for increasing $t_o$, *i. e.* evidence for an effect, and decreasing relative between study heterogeneity $d$. Furthermore, for small $t_o$, increasing $d$ increases the required $c$ much stronger than for large $t_o$. Comparing the shrinkage to the predictive model, the required $c$ under the shrinkage model is much larger for the same $t_o$, especially for small $t_o$. These results illustrate the fact that to achieve a reasonable power, the sample size of the replication study needs to be massively increased compared to the original study, when the results were only suggestive and possibly inflated and/or subject to heterogeneity.

### 2.2.4 Specification of heterogeneity parameter

When taking into account between study heterogeneity, one needs to specify a value for the heterogeneity parameter $\tau^2$ to be able to compute predictions of $\hat{\theta}_r$. However, in the current setting it is not possible to estimate $\tau^2$ using only data from the original study, since $\theta$ in the marginal likelihood of $\hat{\theta}_o \,|\, \theta \sim \mathrm{N}(\theta, \sigma_o^2 + \tau^2)$ is unknown. Therefore, a different method for specifying $\tau^2$ is needed.

Based on the hierarchical model of effect sizes

$$\theta_k \,|\, \theta \sim \mathrm{N}(\theta, \tau^2),$$

95% of the effect sizes $\theta_k$ should lie within the interval $\theta \pm z_{0.025} \cdot \tau$. For the setting where $\theta$ is a log(odds ratio), Spiegelhalter *et al.* (2004) proposed to look at the ratio of the upper and lower limit of this interval on the odds ratio scale. Namely, the ratio of the odds ratio quantiles is $\exp(\theta_{k,97.5\%})/\exp(\theta_{k,0.25\%}) \approx \exp(3.92 \cdot \tau)$. Spiegelhalter *et al.* (2004) argue that it is unlikely for the odds ratio quantiles to vary more than an order of magnitude, *i. e.* $\exp(3.92 \cdot \tau) > 10$ and derived from this a classification, which was also used as a guideline in Neuenschwander *et al.* (2018) for hazard ratio effect sizes.

However, if the effect size $\theta$ is not a log(odds ratio) but a Fisher $z$-transformed correlation $\tanh^{-1}(\rho)$ and assuming $\theta = 0$, then the ratio $\tanh(1.96 \cdot \tau)/\tanh(-1.96 \cdot \tau) = -1$ for all $\tau$, since $\tanh(-x) = -\tanh(x)$. Furthermore, this ratio is not well defined for all combinations of quantiles, because $\tanh(0) = 0$. A more sensible approach for effect sizes which are on the correlation scale is to look at the difference of the quantiles instead of the ratio of quantiles. Because correlations are bounded to the interval between minus one and one, the difference is also bounded, everywhere defined, and easy to interpret. Hence, one can determine which value of $\tau$ leads to the difference of the backtransformed correlation, *i. e.* $\delta = \tanh(\theta_{k,97.5\%}) - \tanh(\theta_{k,2.5\%})$, having a plausible value. Figure 2.8 shows the required heterogeneity $\tau$ as a function of $\delta$ and assuming $\theta = 0$.

However, this raises the question of how one should classify these differences and which value should be picked for the current setting, since the classification by Spiegelhalter *et al.* (2004) was not derived for differences of correlations. In the context of specifying a target effect size for sample size calculations, Cohen (1992) proposed a classification for the magnitude of effect sizes, such as standardized mean differences or correlation coefficients. That is, a medium effect size should reflect an effect which is "visible to the eye", a small effect size should be smaller but not trivial, and finally a large effect size should have the same difference to the medium effect size as the small effect size, but in the other direction. In the setting of direct replication studies, it is reasonable to assume that the between study heterogeneity should not be very large, because these kind of studies are usually matched as closely as possible to the original studies. This suggests a $\tau = 0.08$ leading to $\delta$ being of the size of a medium effect to be a sensible choice.

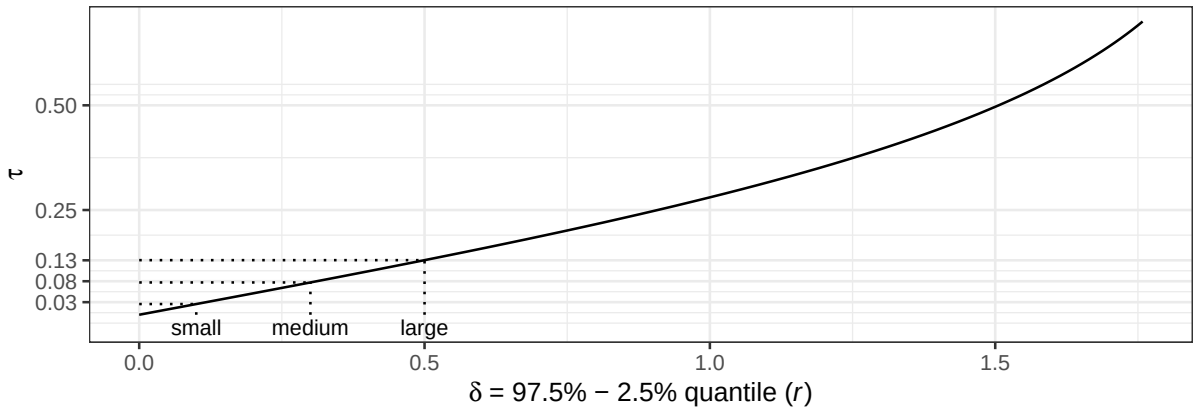


**Figure 2.8:** Between study heterogeneity $\tau$ as a function of $\delta = \tanh(\theta_{k,97.5\%}) - \tanh(\theta_{k,2.5\%})$, difference between quantiles of backtransformed correlations, assuming that $\theta = 0$. The values corresponding to small, medium, and large effect sizes on correlation scale according to the classification by Cohen (1992) are depicted by dotted lines.

However, since this decision is only motivated theoretically, it is advisable to conduct a sensitivity analysis to investigate how much the results would change when choosing different values.

For exploratory purposes it is also possible to estimate $\tau^2$ ad hoc by using the effect estimate of the replication study in addition to the effect estimate of the original study. For instance, one can maximize the likelihood of the predictive distribution when using a flat prior, $\hat{\theta}_r \,|\, \hat{\theta}_o \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2 + \sigma_r^2 + 2\tau^2)$, which leads to a estimator of $\tau^2$ being

$$\hat{\tau}^2 = \max\left\{\frac{(\hat{\theta}_r - \hat{\theta}_o)^2 - \sigma_o^2 - \sigma_r^2}{2}, 0\right\}.$$

On the other hand, when using a sceptical prior there is no analytical expression, but $\hat{\tau}^2$ needs to be obtained numerically. Moreover, instead of the likelihood also other objective functions, such as scoring rules (discussed in the next section), which are more suited for prediction problems can be used to estimate $\tau^2$.

## 2.3 Predictive evaluation methods

To assess the quality of probabilistic predictions, extensive methodology has been developed, see for example Gneiting and Katzfuss (2014). In the following section, various methods that will be used to evaluate the predictions of the replication studies are discussed. The R code of the implemented evaluation methods can be found in Appendix A.1.

### 2.3.1 Discrimination, calibration, and sharpness

When comparing the actual observed events with their predictive distributions, one can distinguish different aspects of this comparison. *Discrimination* characterizes how well a model is able to predict different observations with different predictions. *Calibration*, on the other hand, describes the statistical agreement of the whole predictive distribution with the actual observations, *i. e.* they should be indistinguishable from randomly generated samples from the predictive distribution. One can assess further the *sharpness* aspect of the predictions, *i. e.* the concentration of the predictive distribution. Under the paradigm of *maximizing the sharpness subject to calibration*, for the same calibration, a predictive distribution with smaller variance should be preferred (Gneiting *et al.*, 2007).

### Probability integral transform

A common tool to assess the calibration of continuous predictions is the *probability integral transform* (PIT) which is the value of the predictive cumulative distribution function $F(y) = \mathsf{Pr}(Y \leq y)$ evaluated at the actual observed value $y_o$

$$\text{PIT}(y_o) = F(y_o).$$

If the distribution of the realizations matches the predictive distribution, *i. e.* $Y_o \sim F$, then $Y_o = F^{-1}(U)$, where $U \sim U(0, 1)$, and therefore $F(Y_o) \sim U(0, 1)$. This result implies that the PIT values of well calibrated predictions should follow a standard uniform distribution which is usually assessed visually by examining a histogram of the PIT values (Held and Sabanés Bové, 2014). If the PIT histogram looks U-shaped, the predictive distribution is likely to be underdispersed, while hump-shaped histograms indicated overdispersed predictive distributions. Uniformity of the PIT values can also be assessed using formal tests, *e. g.* a Kolmogorov-Smirnov test for predictions with no dependence structure. Uniform PIT values are a necessary condition for predictions to be well calibrated, however, additional methods should be used to also assess the sharpness of the predictions (Gneiting *et al.*, 2007).

### Area under the curve

In the case of binary outcomes, *area under the curve* (AUC) is commonly used to assess probabilistic predictions regarding their discriminative quality. Let $i$ denote a randomly chosen event with prediction probability $\mathsf{Pr}(Y_i = 1) = \pi_i$, which did actually occur ($y_i = 1$) and let $j$ denote another randomly chosen event with prediction probability $\mathsf{Pr}(Y_j = 1) = \pi_j$, which did not occur ($y_j = 0$), then

$$\text{AUC} = \mathsf{Pr}(\pi_i > \pi_j).$$

The AUC can be estimated in different ways, for instance by numerically integrating the empirical receiver operating characteristic (ROC) curve or by dividing the Wilcoxon rank sum statistic by the product of the number of events and the number of non-events. An AUC of 0.5 is obtained by just randomly guessing the outcome of the event to be predicted. Therefore, only values above 0.5 indicate better than random discrimination. However, predictions which lead to AUC $< 0.5$ can just be inverted to obtain $1 - \text{AUC}$. Confidence intervals for the AUC can be computed in various way, for instance, Wald-type confidence interval can be constructed on original or logit scale (Held and Sabanés Bové, 2014).

### Calibration slope

Originally proposed by Cox (1958), the *calibration slope* method can be used to assess the calibration of predictions by regressing the actual realizations on their predictions, *i. e.* for continuous predictions $y_o = \alpha + \beta\hat{y}_o$ and for binary predictions $\text{logit}(\pi_o) = \alpha + \beta\,\text{logit}(\hat{\pi}_o)$, where $\pi_o = \mathsf{Pr}(y_o = 1)$. A well calibrated prediction model should lead to $\beta \approx 1$, whereas $\beta > 1$ and $\beta < 1$ indicate under- and overestimation respectively. To analyze whether the calibration slope estimate differs statistically significantly from one, standard inferential methods for regression models can be used. Moreover, the estimated calibration slope can also be used as a shrinkage factor to calibrate a model for future use (Steyerberg, 2009).

## 2.3.2 Scoring rules

A *scoring rule* $S(f(y), y_o)$ assigns a real number to density or probability mass function $f(y)$ of a predictive distribution $F(y)$ and the realization $y_o$. Usually, scoring rules are negatively oriented, *i. e.* smaller values of $S(f(y), y_o)$ indicate better predictive performance. Scoring rules

are typically used to assess the performance of a predictive distribution with respect to calibration and sharpness simultaneously. Moreover, a more unusual use case of scoring rules is parameter estimation, *i. e.* if one wants to fit a parametric model $F_\theta(y)$, the *optimum score estimator*

$$\hat{\theta}_{\text{OS}} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} S(f_\theta(y), y_i),$$

provides a principled way to estimate $\theta$ based on a sample $Y_1, \ldots, Y_n$ (Gneiting *et al.*, 2007).

**Proper scoring rules**

A scoring rule $S(f(y), y_o)$ is *proper* if the expected score with respect to the true data generating distribution $Y_o \sim F_o$ is minimized when the predictive distribution is equal to the data generating distribution $F_o$, *i. e.*

$$\mathsf{E}_{f_o}[S(f_o(y), y_o)] \leq \mathsf{E}_{f_o}[S(f(y), y_o)]$$

for all $f(y)$. The scoring rule is *strictly proper* if this holds with equality only if $f(y) = f_o(y)$ (Gneiting *et al.*, 2007).

**Logarithmic score** The *logarithmic score* (LS) is defined as

$$\text{LS}\left(f(y), y_o\right) = -\log f(y_o)$$

and is strictly proper for binary and continuous predictions. In the case of normally distributed predictions $Y \sim \text{N}(\mu, \sigma^2)$, the LS becomes

$$\text{LS}\left(f(y), y_o\right) = \frac{(y_o - \mu)^2}{2\sigma^2} + \log \sigma + \frac{1}{2} \log(2\pi).$$

Furthermore, optimum score estimation based on the logarithmic score leads to maximum likelihood estimation.

**Quadratic score** The *quadratic score* (QS) for continuous predictions is given by

$$\text{QS}\left(f(y), y_o\right) = -2f(y_o) + \int f(t)^2 dt$$

and is strictly proper. For normally distributed predictions $Y \sim \text{N}(\mu, \sigma^2)$, the QS reduces to

$$\text{QS}\left(f(y), y_o\right) = -\frac{2}{\sigma} \varphi\left(\frac{y_o - \mu}{\sigma}\right) + \frac{1}{2\sqrt{\pi}\sigma}.$$

**Continuous ranked probability score** The *continuous ranked probability score* (CRPS) is defined as

$$\begin{aligned} \text{CRPS}\left(f(y), y_o\right) &= \int [F(t) - \mathsf{I}_{[y_o, \infty)}(t)]^2 dt \\ &= \mathsf{E}_F\left\{|Y_1 - y_o|\right\} - \frac{1}{2} \mathsf{E}_F\left\{|Y_1 - Y_2|\right\}, \end{aligned}$$

where $Y_1$ and $Y_2$ are independent random variables with cumulative distribution function $F$. The CRPS has many attractive properties, *e. g.* the CRPS can be used to compare point predictions to probabilistic predictions because it is a generalization of the absolute error. Furthermore, the CRPS uses the same units as the observations. For normally distributed predictions $Y \sim \text{N}(\mu, \sigma^2)$, the CRPS becomes

$$\text{CRPS}\left(f(y), y_o\right) = \sigma \left[\frac{y_o - \mu}{\sigma} \left\{2\Phi\left(\frac{y_o - \mu}{\sigma}\right) - 1\right\} + 2\varphi\left(\frac{y_o - \mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right].$$

**Dawid-Sebastiani score** The *Dawid-Sebastiani score* (DSS) is a scoring rule which only depends on the first two central moments of a predictive distribution, $\mu_{f(y)}$ and $\sigma^2_{f(y)}$. The DSS is given by

$$\mathrm{DSS}\left(f(y), y_o\right) = \frac{(y_o - \mu_{f(y)})^2}{\sigma^2_{f(y)}} + 2\log\sigma^2_{f(y)}$$

and is also strictly proper. In the case of normally distributed predictions $Y \sim \mathrm{N}(\mu, \sigma^2)$, the DSS is the same as the logarithmic score up to an affine transformation (Gneiting and Katzfuss, 2014).

**Brier score** The *Brier score* (BS) is a scoring rule specific to binary predictions. Denote the predictive distribution $f(y) = \hat{\pi}$ for $y = 1$ and as $f(y) = 1 - \hat{\pi}$ for $y = 0$, then the BS is defined as

$$\mathrm{BS}\left(f(y), y_o\right) = (y_o - \hat{\pi})^2,$$

and is strictly proper. To allow model comparison on data with different prevalences of events, often $\overline{\mathrm{BS}}$, the mean BS of a set of predictions, is normalized by

$$\overline{\mathrm{BS}}^* = (\overline{\mathrm{BS}}_0 - \overline{\mathrm{BS}})/\overline{\mathrm{BS}}_0,$$

where $\overline{\mathrm{BS}}_0 = \sum_{i=1}^n (y_i - \bar{y})^2/n = \bar{y}(1-\bar{y})$, is the BS of the prevalence prediction $\bar{y} = n^{-1}\sum_{i=1}^n y_i$. Moreover, $\overline{\mathrm{BS}}_0$ also serves as an upper bound for useful predictions (Held and Sabanés Bové, 2014).

### Score based miscalibration tests

Calibration of predictive distributions can also be assessed using formal significance tests based on the observed scores.

Spiegelhalter (1986) proposed a test based on the Brier score, which is today known as *Spiegelhalter's z-test*. First, note that the mean Brier score can be decomposed into

$$\overline{\mathrm{BS}} = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\pi}_i)(1 - 2\hat{\pi}_i) + \frac{1}{n}\sum_{i=1}^n \hat{\pi}_i(1 - \hat{\pi}_i),$$

where the first term measures calibration and the second term measures sharpness. Under the null hypothesis of perfect calibration, *i. e.* $\mathsf{E}(y_i) = \hat{\pi}_i$, the expectation of the first term is zero and thus the second term corresponds to $\mathsf{E}\left(\overline{\mathrm{BS}}\right)$. The variance of the mean Brier score under the null hypothesis can be derived to be

$$\mathrm{Var}(\overline{\mathrm{BS}}) = \frac{1}{n^2}\sum_{i=1}^n (1 - 2\hat{\pi}_i)^2\hat{\pi}_i(1 - \hat{\pi}_i).$$

Hence $z_{\mathrm{BS}} = \left\{\overline{\mathrm{BS}} - \mathsf{E}\left(\overline{\mathrm{BS}}\right)\right\} / \mathrm{Var}(\overline{\mathrm{BS}})^{1/2}$ is approximately standard normal distributed under perfect calibration and can be used as a miscalibration test.

Held *et al.* (2010) proposed similar tests to assess the calibration of continuous predictions. For normally distributed predictions under the null hypothesis of perfect calibration, expectation and variance of the mean logarithmic and CRP scores can be derived to be

$$\mathsf{E}\left(\overline{\mathrm{LS}}\right) = \frac{1}{2} + \frac{1}{n}\sum_{i=1}^n \log\sigma_i + \frac{1}{2}\log(2\pi) \qquad \mathrm{Var}\left(\overline{\mathrm{LS}}\right) = \frac{1}{2n}$$

$$\mathsf{E}\left(\overline{\mathrm{CRPS}}\right) = \frac{1}{\sqrt{\pi}}\frac{1}{n}\sum_{i=1}^n \sigma_i \qquad \mathrm{Var}\left(\overline{\mathrm{CRPS}}\right) = \frac{C}{n^2}\sum_{i=1}^n \sigma_i^2,$$

where $C = 1/3 - (4 - \sqrt{12})/\pi \approx 0.16275$. Using these results, the test statistics $z_{\text{LS}} = \left\{\overline{\text{LS}} - \mathsf{E}\left(\overline{\text{LS}}\right)\right\} / \text{Var}(\overline{\text{LS}})^{1/2}$ and $z_{\text{CRPS}} = \left\{\overline{\text{CRPS}} - \mathsf{E}\left(\overline{\text{CRPS}}\right)\right\} / \text{Var}(\overline{\text{CRPS}})^{1/2}$ can be constructed. Under the null hypothesis of perfect calibration both follow asymptotically a standard normal distribution and can therefore be used to test for miscalibration.

Moreover, Held *et al.* (2010) also proposed score based miscalibration tests using regression models. By conditioning on characteristics of the predictive distribution, these approaches can provide more powerful tools to detect miscalibration compared to unconditional tests. For a perfectly calibrated prediction $f(y)$, the expected DS score is $\mathsf{E}\left\{\text{DSS}(f(y), y_o)\right\} = 1 + 2\log\sigma$ with variance $\text{Var}\left\{\text{DSS}(f(y), y_o)\right\} = 1$. Using these results, a regression model can be formulated, namely

$$\text{DSS}_i = a + b\log\sigma_i + \epsilon_i,$$

were $\epsilon_i$ are independent errors with zero mean. Under the null hypothesis of perfect calibration, $a = a_0 = 1$ and $b = b_0 = 2$. Hence, from the least-squares fit with coefficients $\hat{a}, \hat{b}$ and estimated variance-covariance matrix $\hat{V}$, the test statistic

$$T_{\text{DSS}} = (\hat{a} - a_0, \hat{b} - b_0)\hat{V}^{-1}(\hat{a} - a_0, \hat{b} - b_0)^T$$

follows asymptotically a $\chi^2$-distribution with two degrees of freedom. Using a similar approach, one can define a regression model for the CRP scores

$$\text{CRPS}_i = c + d\sigma_i + \epsilon_i.$$

Since $\text{Var}(\text{CRPS}) \propto \sigma_i^2$, a heteroscedastic model with weights $1/\sigma_i^2$ should be used. Under the null hypothesis of perfect calibration and assuming normality, $c = c_0 = 0$ and $d = d_0 = 1/\sqrt{\pi}$. A test statistic can be constructed in in the same way as in the DSS-regression case.

## 2.4 Data

Several data sets were used to compare the different prediction methods. In all data sets, effect estimates were provided on the correlation scale. If not already present, the Fisher $z$-transformation was applied to the effect estimates and the corresponding standard errors were computed. All R code for data preprocessing can be found in Appendix A.2.

### Reproducibility Project Psychology

Open Science Collaboration (2015) conducted 100 replications of studies from the field of psychology. The sampling frame of the original studies was chosen to minimize potential selection bias and maximize generalizability of the findings. Namely, the sampling frame consisted of all articles published in the journals *Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition* within the year 2008. By default, the last experiment of a study was picked to be replicated and for each of these experiments a key statistical inference test, such as a $t$-test or $F$-test, was selected as the focus of the statistical comparison.

All files were downloaded from `https://github.com/CenterForOpenScience/rpp/archive/master.zip`. The `masterscript.R` was run and the data then taken from the generated `MASTER` object. The standard errors of the Fisher $z$-transformed correlation coefficients were obtained by binding the `final$sei.o` and `final$sei.r` vectors with the remaining data. According to the supplementary material, the $p$-values from the studies with ID's 7, 15, 47, 94, 120, and 140 were one-sided and were therefore multiplied by two to obtain two-sided $p$-values. Only the "meta-analytic subset" was used, which consists of 73 studies where the standard error of the Fisher $z$-transformed effect estimates can be computed.

### Experimental Economics Replication Project

Camerer *et al.* (2016) conducted 18 replications of studies from the field of experimental economics. The sampling frame consisted of all studies published between 2011 and 2015 in the journals *American Economic Review* and *Quarterly Journal of Economics*, which reported at least one statistically significant between subject treatment effect. If more than one statistically significant treatment effect was reported, "the most central result" based on the extent of emphasis in the publication was chosen as the key statistical result. In the case of more than one central results, Camerer *et al.* picked the result that was the most efficient to replicate, which they justify by "efficiency is central to economics". When there was still ambiguity, the procedure of the reproducibility project psychology to pick the last result, was chosen.

For this replication project also a *prediction market* was conducted in order to estimate the peer beliefs about whether a replication study will result in a statistically significant result. Prediction markets are a tool to aggregate beliefs of market participants about the possibility of an investigated outcome and they have been used successfully in numerous domains, *e. g.* sports or politics (Dreber *et al.*, 2015). Since the estimated peer beliefs are also probabilistic predictions, they can be compared to the probability of significance under the discussed statistical prediction methods.

All files were downloaded from `https://osf.io/pnwuz/`. However, to "generate" the data from the file `create_studydetails.do`, the commercial software `STATA` is required. Since the data set is very small, the required data were manually extracted from the code in the file `create_studydetails.do`. To compute the standard errors of the Fisher $z$-transformed effect estimates, the sample sizes reported in the `effectdata.py` file rather than the ones reported in the `create_studydetails.do` were taken. The former correspond to the effective sample sizes while the latter in some cases corresponds to the number of measurements, which lead to different prediction intervals than the ones reported in the publication (however, in all tables Camerer *et al.* (2016) report the larger "number of measurements" sample size). The data regarding the prediction market and survey beliefs were also manually extracted from table S3 in the supplementary material, which was downloaded from `http://science.sciencemag.org/content/suppl/2016/03/02/science.aaf0918.DC1`.

### Social Sciences Replication Project

Camerer *et al.* (2018) conducted 21 replications of studies from the field of social sciences. The sampling frame consisted of all social sciences studies published in the journals *Nature* and *Science* between 2010 and 2015. Furthermore, the studies needed to have either a within or between subjects treatment comparison design and the included experiments had to be performed in a standard lab using student subjects or other easily accessible adult subjects. Finally, there had to be at least one statistically significant finding reported in the studies. The treatment effect to be replicated was by default selected as the first experiment reported in the publications which achieved statistical significance. In this replication project a slightly different approach was used in the conduct of the replication studies. In a first stage, the replication studies had 90% power to detect 75% of the original effect estimate at $\alpha = 0.05$ with a two-sided test. If statistical significance of the test result was not obtained, a second data collection was carried out with power of 90% to detect 50% of the original effect estimate. The data of both data collections were then pooled together.

Similarly 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 can be used as a comparison to the statistical predictions.

The data were taken from the `D3 - ReplicationResults.csv` file, which was downloaded from `https://osf.io/abu7k`. For replications which underwent only the first stage, the data from the first stage were taken as the data for the replication study. For the replications which

reached the second stage, the pooled data from both stages were taken as the data for the replication study. Additionally, the data regarding survey and prediction market beliefs were extracted from the `D6 - MeanPeerBeliefs.csv` file, which was downloaded from https://osf.io/vr6p8/.

### Experimental Philosophy Replicability Project

Cova *et al.* (2018) conducted 40 replications of studies from the field of experimental philosophy. The sampling frame consisted of all studies between 2003 and 2015 which were listed on the experimental philosophy page of the university of Yale and which were as well published in one of 35 journals in which experimental philosophy research is usually published (a list defined by the coordinators of this project). For each year between 2003 and 2015, three studies were selected, one of them being the most cited of the year, while the other two were randomly selected. By default the first experiment of a publication was selected to be replicated.

All data were taken from the `XPhiReplicability_CompleteData.csv` file, which was downloaded from https://osf.io/4ewkh. However, only a subset of 31 of these replications could be used, since only for these data, effect estimates on correlation scale and effective sample size for original and replication were available simultaneously. Because $p$-values were most of the time reported as inequalities, they were recalculated using a normal approximation on the Fisher $z$ scale.

## 2.5 Software

All analyses were performed in the R programming language (R Core Team, 2019) using base packages and the following analysis-specific packages: Packages from the `tidyverse` were used for data preparation and plotting (Wickham, 2017). The formatting of the $p$-values as well as the computation of the AUCs with confidence intervals were performed using the `biostatUZH` package which is available on http://ebuzh.r-forge.r-project.org/. The nested tables were generated using the `tables` package (Murdoch, 2018).

# Chapter 3

# Results

## 3.1 Descriptive results

Figure 3.1 shows plots of the original vs. the replication effect estimate, both on the correlation scale. Most effect estimates of the replication studies are considerably smaller than those of

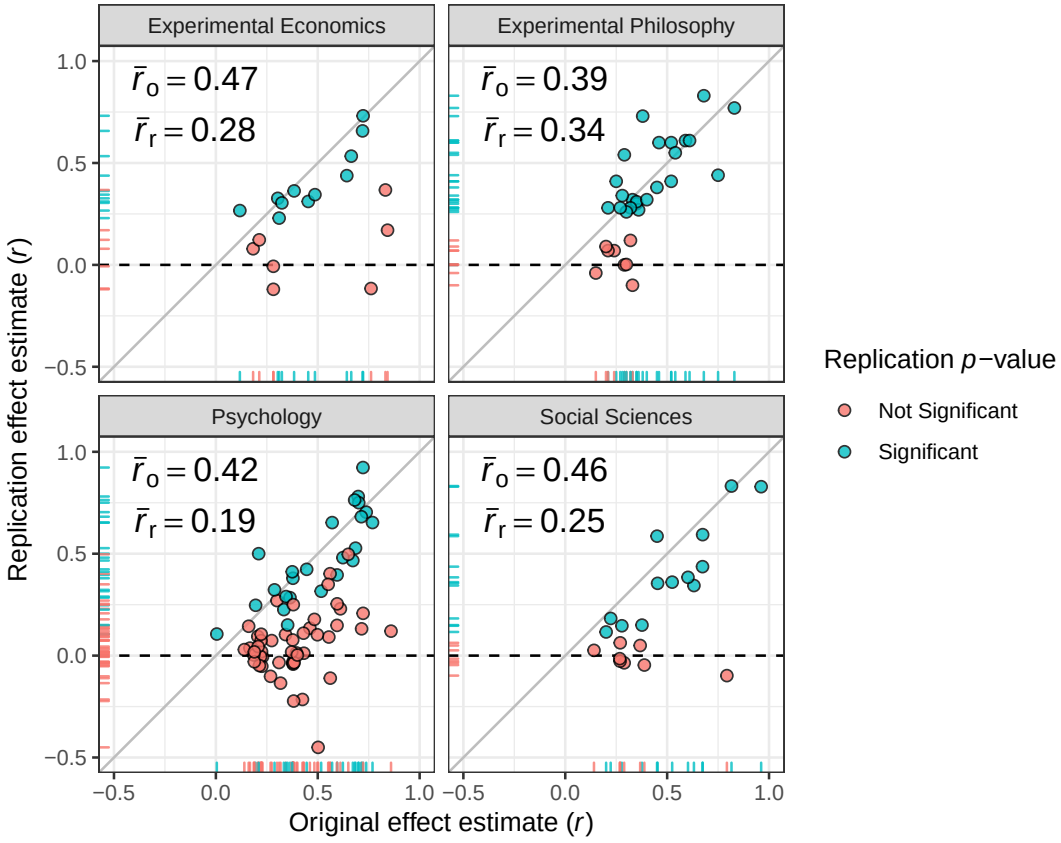


**Figure 3.1:** Effect estimate of original study vs. effect estimate of replication study (on correlation scale). The color of the points indicates whether statistical significance at the 0.05 level was achieved.

the original studies. Namely, the mean effect estimates of the replications are roughly half as large as the mean effect estimates of the original studies. An exception are the predictions in the philosophy data set, where the mean effect estimate only decreased from 0.39 to 0.34. Furthermore, studies showing a comparable effect estimate in the replication and original study usually also achieved statistical significance, while studies showing a large decrease of the effect

estimate were less likely to achieve statistical significance in the replication.

In Figure 3.2 a comparison of the original and the replication study $p$-values is shown. On average, the $p$-values of the original studies are much smaller than the $p$-values of the

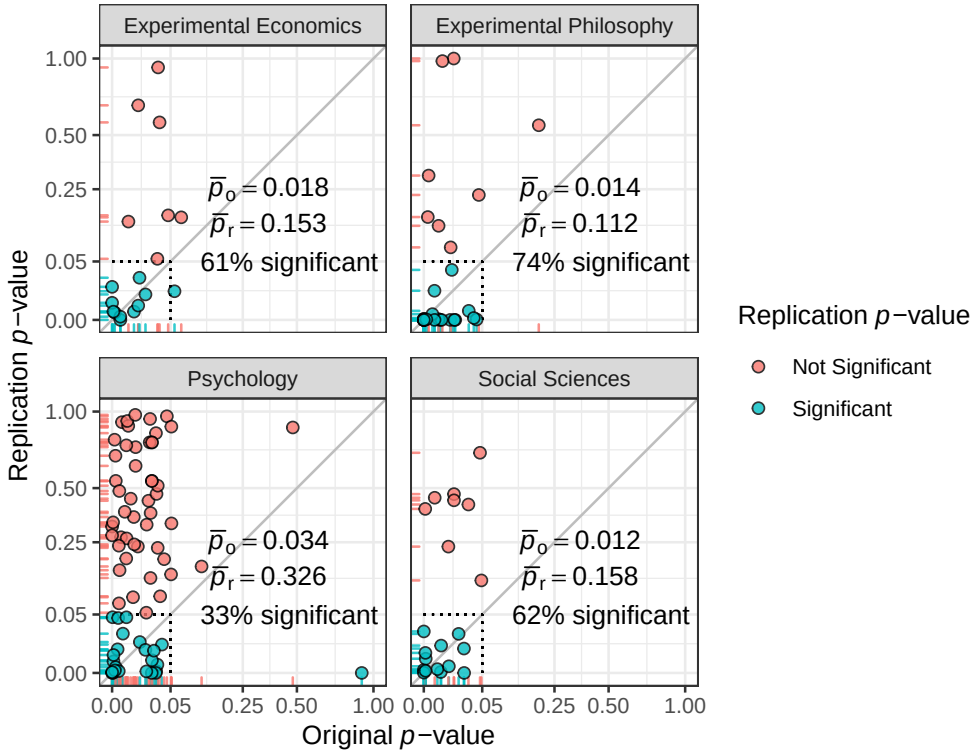


**Figure 3.2:** $p$-value of original study vs. $p$-value of replication study.

replication studies. Moreover, many of the replication studies show $p$-values above the threshold 0.05, whereas most original studies showed a $p$-value smaller than 0.05. In the psychology data set only 33% of the replications achieve statistical significance, while in the social sciences and economics data sets around 60% of the replications show significant effect estimates. Finally, in the philosophy data set 74% of the replication studies achieve statistical significance.

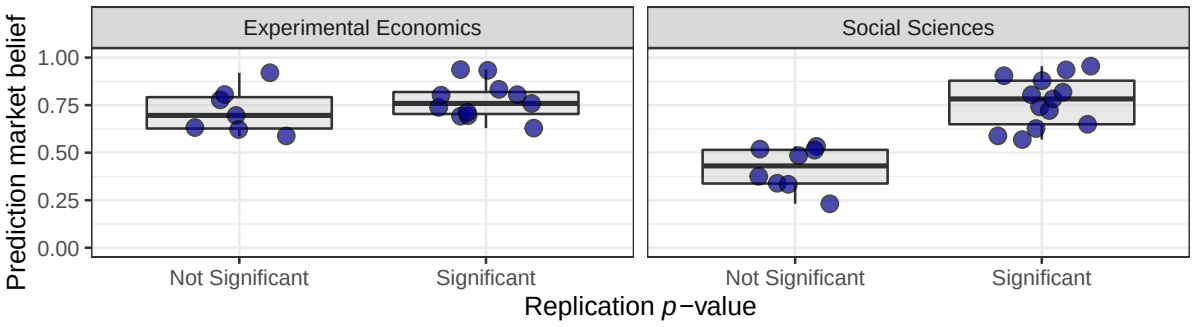


**Figure 3.3:** Statistical significance of replication study vs. estimated prediction market beliefs about whether the replication studies will achieve statistical significance (at $\alpha = 0.05$).

Figure 3.3 illustrates the elicited prediction market beliefs about whether the replication studies will achieve statistical significance. In the case of the social science data set, the elicited beliefs show a perfect separation with respect to the statistical significance of the replication

study. On the other hand, the distribution of the prediction market beliefs in the economics data set is very similar for significant and non-significant replications.

## 3.2 Predictive evaluation

### 3.2.1 Continuous predictions

In the following section the predictive performance of the investigated prediction methods will be evaluated with methods suited for continuous predictive distributions.

#### Prediction intervals

Figure 3.4 shows again plots of the original vs. the replication effect estimates. In addition, the corresponding 95% prediction interval is vertically shown around each study pair. Comparing

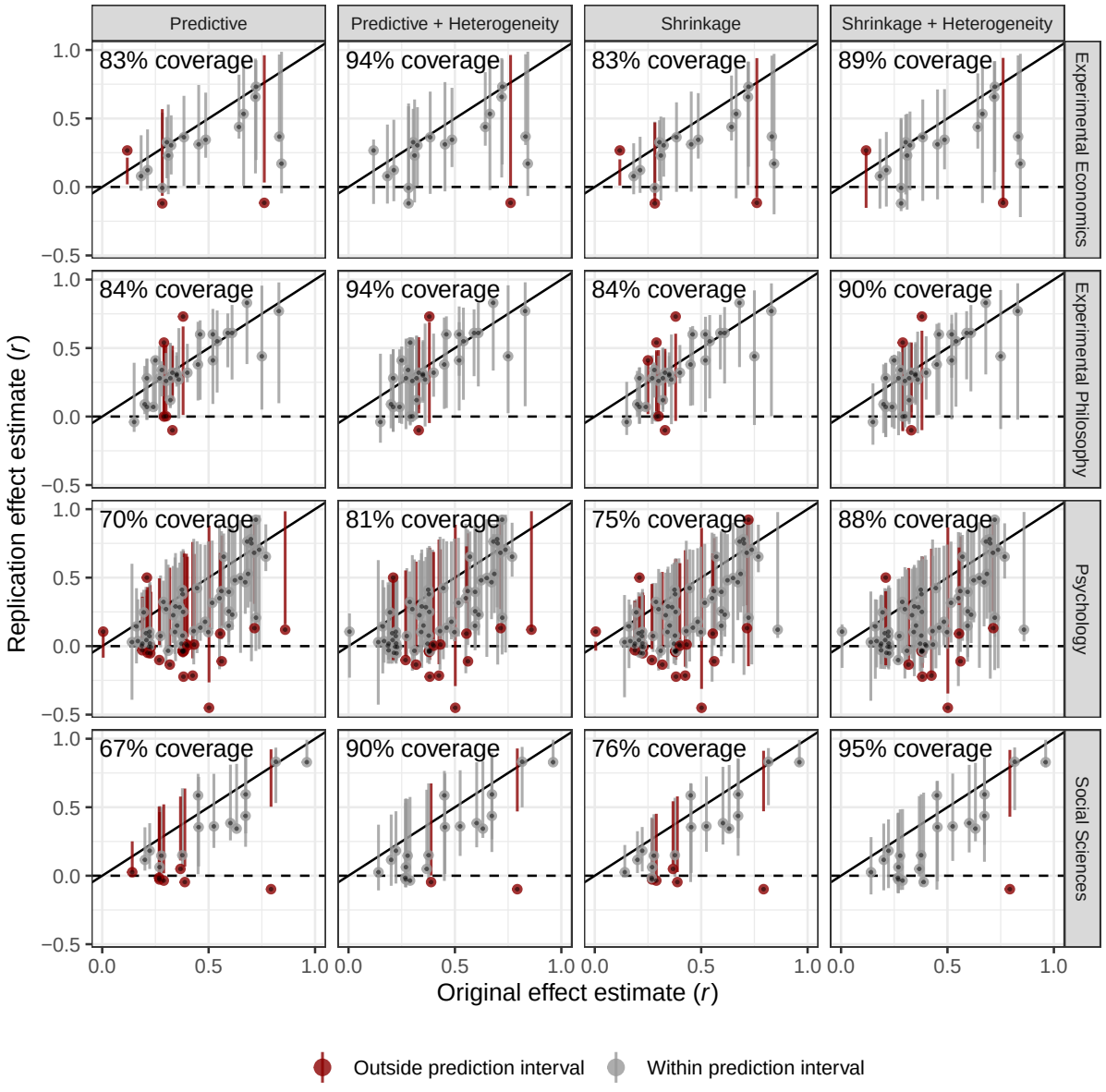


**Figure 3.4:** 95% prediction intervals of the replication effect estimates.

the different methods across data sets, the shrinkage method shows the same coverage as the

predictive method in the economics and philosophy data sets, whereas in the psychology and social sciences data sets the shrinkage method shows a higher coverage compared to the predictive method. As is to be expected, when taking into account heterogeneity, the prediction intervals become wider and the coverage improves considerably in all cases. In the philosophy and economics data sets the highest coverage is achieved for the non-shrunken predictions from the predictive method, while in the psychology and social sciences data sets the highest coverage is achieved for the shrinkage predictions. Moreover, in all data sets except the psychology data set, the best method is able to achieve nominal coverage of about 95%, whereas in the psychology data set the best method achieves slightly less. These improvements in coverage suggest improved calibration of the predictions taking into account heterogeneity (and shrinkage in the case of the social sciences and psychology data sets). Finally, in the psychology and social sciences data sets, the replication effect estimates that are not covered by their prediction intervals are usually smaller than the lower limits of the intervals. In the economics and philosophy data sets, on the other hand, the non-coverage appears to be more symmetric.

### Predicted means

Table 3.1 shows a comparison of the predicted and the actually observed mean of the replication effect estimates on the correlation scale. Since the predicted mean is the same for both predictive methods, only one number is shown. As it is to be expected, the predicted mean effect estimate

**Table 3.1:** Observed vs. predicted mean replication effect estimates on correlation scale.

| Project | Method | mean($r_r$) | mean($\hat{r}_r$) |
|---|---|---|---|
| Experimental Economics | Predictive | 0.28 | 0.47 |
| $n = 18$ | Shrinkage | 0.28 | 0.42 |
| | Shrinkage and Heterogeneity | 0.28 | 0.41 |
| Experimental Philosophy | Predictive | 0.34 | 0.39 |
| $n = 31$ | Shrinkage | 0.34 | 0.35 |
| | Shrinkage and Heterogeneity | 0.34 | 0.34 |
| Psychology | Predictive | 0.19 | 0.42 |
| $n = 73$ | Shrinkage | 0.19 | 0.37 |
| | Shrinkage and Heterogeneity | 0.19 | 0.36 |
| Social Sciences | Predictive | 0.25 | 0.46 |
| $n = 21$ | Shrinkage | 0.25 | 0.42 |
| | Shrinkage and Heterogeneity | 0.25 | 0.41 |

under the shrinkage method is smaller than the predicted mean effect estimate under the predictive method (which is just the mean effect estimate of the original studies). Furthermore, the shrinkage method taking into account heterogeneity shows the smallest predicted mean effect estimate. However, with exception of the philosophy data set, the predicted mean effect estimate is still substantially larger than the observed mean effect estimate even for the shrinkage method taking heterogeneity into account, suggesting overestimation.

### Calibration slope

In Figure 3.5 the calibration slopes obtained by regressing the Fisher $z$-transformed replication effect estimates on the mean parameter of their predictive distributions are shown with 95% confidence intervals. Usually, within one data set the slopes do not differ very much between the four methods. Comparing the different data sets, the social sciences and psychology data sets show calibration slopes of around 0.7, while the calibration slopes in the economics data set

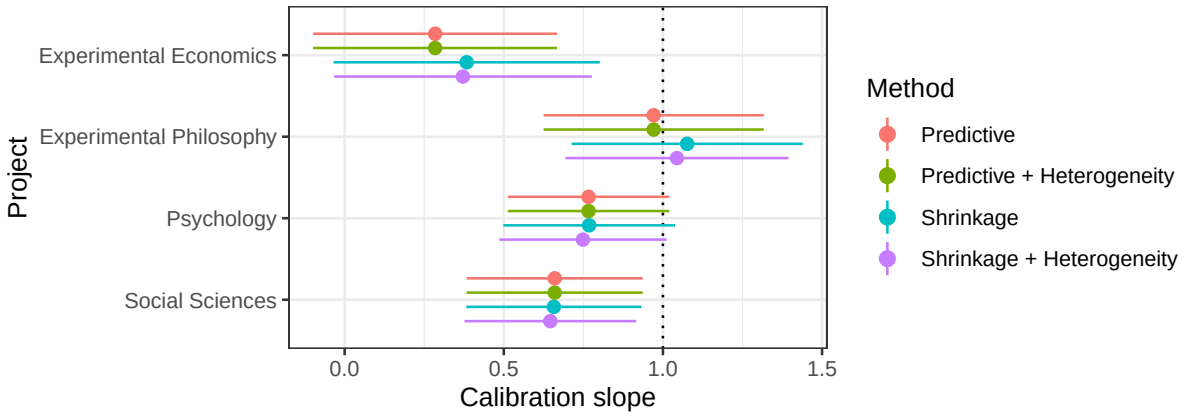


**Figure 3.5:** Calibration slope for continuous predictions with 95% confidence interval.

are around 0.4, suggesting overestimation of the mean parameter of all predictive distributions. The slopes in the philosophy data set, on the other hand, show values around one, suggesting that the mean parameters are well calibrated. Moreover, since the sample size in most projects is small, the confidence intervals of the calibration slopes are usually very wide.

### PIT histograms

Figure 3.6 shows histograms of the PIT values of the four prediction methods, where the range from zero to one has been divided equally into eight bins. In some of the histograms in the social

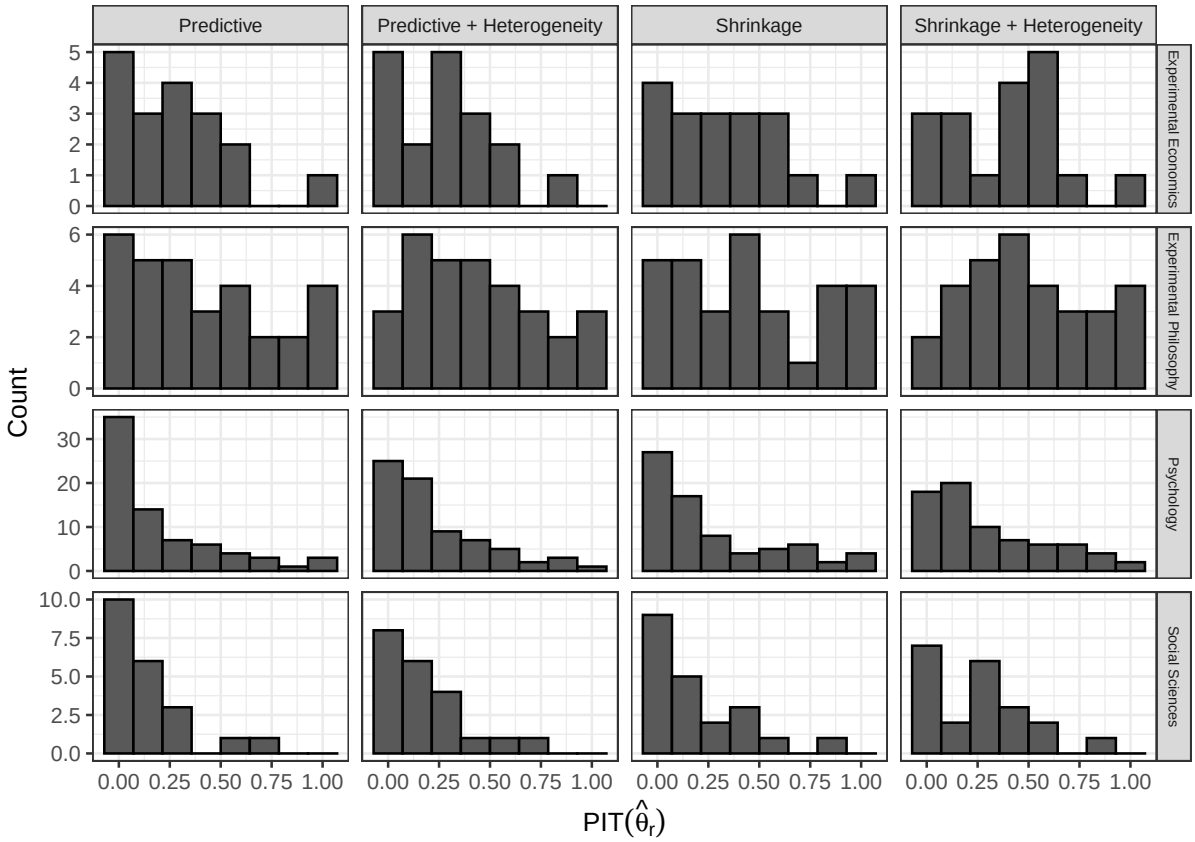


**Figure 3.6:** Histograms of PIT values.

sciences and economics data sets there are bins with zero observations, however, these data sets also have the smallest sample sizes. Comparing the PIT histograms between the different methods, differences in the uniformity of the PIT values are visible. In the psychology and social sciences data sets, the predictive method not taking into account heterogeneity shows extreme bumps in the lower range of the PIT values. On the other hand, the histograms of the predictive method taking into account heterogeneity, and the histograms of both shrinkage methods look flatter, suggesting less miscalibration. In the case of the economics data sets, the PIT histograms of both shrinkage methods look uniform, while the histograms of the predictive methods look more skewed, indicating less miscalibration of the former compared to the latter. Finally, in the philosophy data set the histograms look acceptable for all methods, suggesting no severe miscalibration.

**Table 3.2:** Kolmogorov-Smirnov tests comparing PIT values to $U(0,1)$ distribution.

| Project | Method | Test statistic | $p$-value |
|---|---|---|---|
| Experimental Economics | Predictive | 0.38 | 0.007 |
| $n = 18$ | Predictive and Heterogeneity | 0.39 | 0.006 |
| | Shrinkage | 0.30 | 0.061 |
| | Shrinkage and Heterogeneity | 0.29 | 0.073 |
| Experimental Philosophy | Predictive | 0.21 | 0.11 |
| $n = 31$ | Predictive and Heterogeneity | 0.19 | 0.20 |
| | Shrinkage | 0.18 | 0.26 |
| | Shrinkage and Heterogeneity | 0.08 | 0.97 |
| Psychology | Predictive | 0.48 | $< 0.0001$ |
| $n = 73$ | Predictive and Heterogeneity | 0.44 | $< 0.0001$ |
| | Shrinkage | 0.41 | $< 0.0001$ |
| | Shrinkage and Heterogeneity | 0.36 | $< 0.0001$ |
| Social Sciences | Predictive | 0.61 | $< 0.0001$ |
| $n = 21$ | Predictive and Heterogeneity | 0.54 | $< 0.0001$ |
| | Shrinkage | 0.52 | $< 0.0001$ |
| | Shrinkage and Heterogeneity | 0.42 | 0.0008 |

Table 3.2 shows the results of Kolmogorov-Smirnov tests applied to the PIT values to test for miscalibration. In each data set, the test statistic of the shrinkage method taking into account heterogeneity shows the smallest value, suggesting that there is the least evidence for this method to be miscalibrated. Looking at the economics data set, the tests provide weak evidence for miscalibration of the shrinkage methods and moderate evidence for miscalibration of the predictive methods. In the philosophy data set, on the other hand, there is no evidence for miscalibration of any of the methods. Finally, in the social sciences and psychology data sets, the tests provide substantial evidence for miscalibration of all methods.

### Scoring rules

Table 3.3 shows the mean of the logarithmic scores (LS), continuous ranked probability scores (CRPS), and quadratic scores (QS) for each data set and prediction method. Additionally, the same mean scores, as well as their standard errors are are shown in Figure 3.7. It should be noted, that the standard errors are presented only to illustrate the spread of the individual scores and not to compare the mean scores between the different methods (this will be done further below using a paired test). When comparing the methods, the shrinkage method taking into account heterogeneity shows the lowest mean score across all score types in all data sets, suggesting that this method achieves the best predictive performance. The predictive method not taking into

account heterogeneity, on the other hand, usually showed the highest mean score across all score types, indicating that this method performs the worst among the four methods.

**Table 3.3:** Mean scores for continuous predictions.

| | | Type | | |
|---|---|---|---|---|
| | | LS | CRPS | QS |
| Project | Method | mean | mean | mean |
| Experimental Economics | Predictive | 0.34 | 0.21 | −0.83 |
| $n = 18$ | Predictive and Heterogeneity | 0.18 | 0.21 | −1.14 |
| | Shrinkage | 0.17 | 0.17 | −1.17 |
| | Shrinkage and Heterogeneity | 0.02 | 0.17 | −1.32 |
| Experimental Philosophy | Predictive | −0.05 | 0.12 | −1.33 |
| $n = 31$ | Predictive and Heterogeneity | −0.18 | 0.12 | −1.51 |
| | Shrinkage | −0.06 | 0.12 | −1.46 |
| | Shrinkage and Heterogeneity | −0.20 | 0.11 | −1.67 |
| Psychology | Predictive | 0.87 | 0.22 | −0.07 |
| $n = 73$ | Predictive and Heterogeneity | 0.51 | 0.22 | −0.55 |
| | Shrinkage | 0.86 | 0.19 | −0.15 |
| | Shrinkage and Heterogeneity | 0.27 | 0.18 | −0.85 |
| Social Sciences | Predictive | 0.85 | 0.22 | −0.17 |
| $n = 21$ | Predictive and Heterogeneity | 0.55 | 0.21 | −0.67 |
| | Shrinkage | 0.54 | 0.19 | −0.58 |
| | Shrinkage and Heterogeneity | 0.25 | 0.18 | −1.17 |

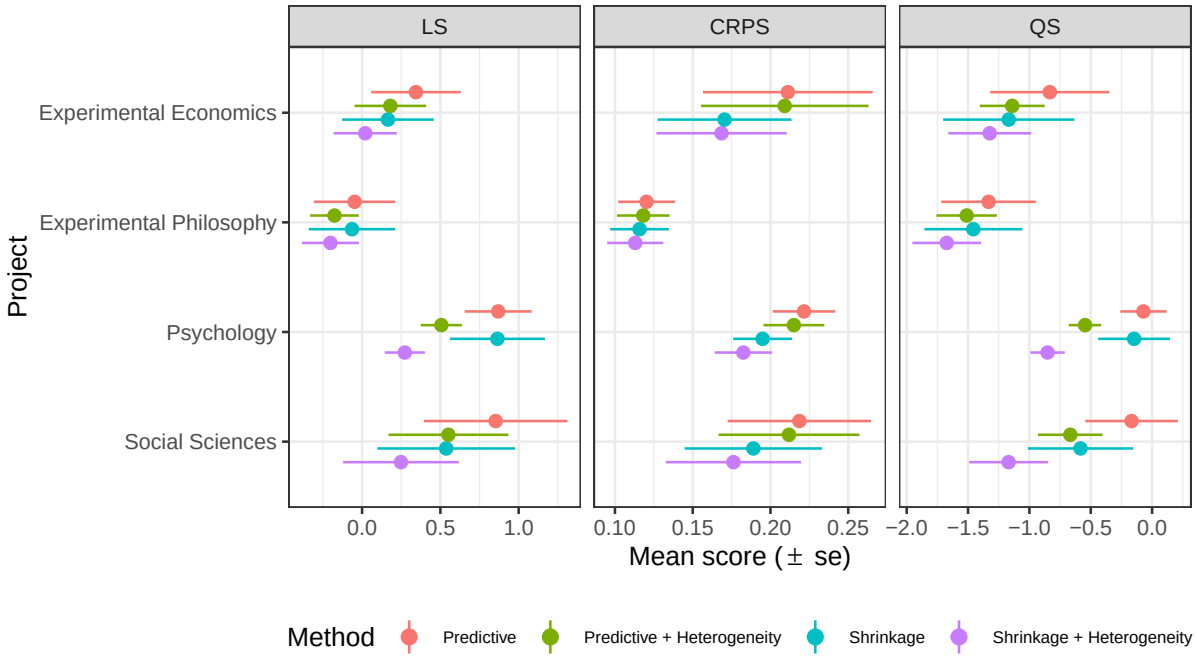


**Figure 3.7:** Mean scores with standard errors.

Table 3.4 shows the $p$-values of the paired Wilcoxon rank sum tests of the scores of the shrinkage method taking into account heterogeneity compared to the scores of the other three prediction methods. Only these comparisons are reported since the shrinkage method taking into

account heterogeneity achieved the lowest mean score in all score types and data sets. For the

**Table 3.4:** Results of paired Wilcoxon rank sum tests of the shrinkage method taking into account heterogeneity vs. the other three methods.

| | | Type | | |
|---|---|---|---|---|
| Project | Test | LS $p$-value | CRPS $p$-value | QS $p$-value |
| Experimental Economics | Predictive | 0.016 | 0.007 | 0.038 |
| $n = 18$ | Predictive and Heterogeneity | 0.005 | 0.003 | 0.014 |
| | Shrinkage | 0.77 | 0.70 | 0.73 |
| Experimental Philosophy | Predictive | 0.95 | 0.66 | 0.44 |
| $n = 31$ | Predictive and Heterogeneity | 0.28 | 0.47 | 0.26 |
| | Shrinkage | 0.79 | 0.95 | 0.36 |
| Psychology | Predictive | $< 0.0001$ | $< 0.0001$ | $< 0.0001$ |
| $n = 73$ | Predictive and Heterogeneity | $< 0.0001$ | $< 0.0001$ | $< 0.0001$ |
| | Shrinkage | $< 0.0001$ | $< 0.0001$ | $< 0.0001$ |
| Social Sciences | Predictive | 0.0007 | 0.0004 | 0.001 |
| $n = 21$ | Predictive and Heterogeneity | $< 0.0001$ | 0.0005 | 0.0003 |
| | Shrinkage | 0.07 | 0.001 | 0.018 |

predictions in the psychology and social sciences data sets, there is in most cases strong evidence of a difference in scores between the shrinkage method taking into account heterogeneity and the other methods. In the philosophy data set, on the other hand, there is no evidence for a difference between the scores of the shrinkage method taking into account heterogeneity and the other methods. Finally, in the economics data set there is moderate evidence for a difference of the scores between the shrinkage method taking into account heterogeneity and the non-shrinkage methods, however, no evidence for a difference of the scores between the two shrinkage methods.

In Table 3.5, the results of the scoring rule based miscalibration tests are shown. The results of the four tests are often but not always in agreement. First, the unconditional test based on the logarithmic score suggests that all methods in the psychology and social sciences data sets are miscalibrated. The test also provides some evidence for miscalibration of the methods which do not take into account heterogeneity in the economics and philosophy data sets. Second, the results from the unconditional test based on the CRPS provide evidence for miscalibration of all methods in the social sciences and psychology data sets, moderate evidence for miscalibration of the predictive method in the economics data set, and weak evidence for miscalibration of the predictive and shrinkage method in the case of the philosophy data set. Third, the DSS regression test indicates miscalibration of all methods in the psychology data set and it provides weak evidence for miscalibration of the predictive method in the social sciences data set. Furthermore, the test does not suggest miscalibration of any method in the economics and philosophy data sets. Finally, the results from the CRPS regression test provide strong evidence for miscalibration of all methods in the psychology data set, no evidence for miscalibration of any method in the philosophy data sets, weak evidence for miscalibration of the methods which do not take into account heterogeneity in the case of the economics data set, and weak evidence for miscalibration of all methods except the shrinkage method taking into account heterogeneity in the social sciences data set.

**Table 3.5:** Results from scoring rule based miscalibration tests.

| | | Test type | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | LS | | CRPS | | DSS-Regression | | CRPS-Regression | |
| Project | Method | Statistic | $p$-value | Statistic | $p$-value | Statistic | $p$-value | Statistic | $p$-value |
| Experimental Economics | Predictive | 2.85 | 0.004 | 2.30 | 0.021 | 3.44 | 0.18 | 6.01 | 0.05 |
| $n = 18$ | Predictive and Heterogeneity | 0.72 | 0.47 | 1.52 | 0.13 | 2.97 | 0.23 | 1.91 | 0.39 |
| | Shrinkage | 2.11 | 0.035 | 1.23 | 0.22 | 3.40 | 0.18 | 7.01 | 0.03 |
| | Shrinkage and Heterogeneity | 0.20 | 0.84 | 0.57 | 0.57 | 0.13 | 0.94 | 0.04 | 0.98 |
| Experimental Philosophy | Predictive | 3.84 | 0.0001 | 2.02 | 0.044 | 3.67 | 0.16 | 3.98 | 0.14 |
| $n = 31$ | Predictive and Heterogeneity | 0.48 | 0.63 | 0.19 | 0.85 | 0.19 | 0.91 | 0.13 | 0.94 |
| | Shrinkage | 4.09 | $< 0.0001$ | 2.13 | 0.033 | 3.55 | 0.17 | 3.79 | 0.15 |
| | Shrinkage and Heterogeneity | 0.80 | 0.42 | 0.32 | 0.75 | 0.33 | 0.85 | 0.36 | 0.83 |
| Psychology | Predictive | 12.86 | $< 0.0001$ | 8.09 | $< 0.0001$ | 31.13 | $< 0.0001$ | 44.76 | $< 0.0001$ |
| $n = 73$ | Predictive and Heterogeneity | 6.25 | $< 0.0001$ | 5.49 | $< 0.0001$ | 16.48 | 0.0003 | 19.82 | $< 0.0001$ |
| | Shrinkage | 13.57 | $< 0.0001$ | 6.68 | $< 0.0001$ | 45.17 | $< 0.0001$ | 80.35 | $< 0.0001$ |
| | Shrinkage and Heterogeneity | 4.28 | $< 0.0001$ | 3.86 | 0.0001 | 8.18 | 0.017 | 9.56 | 0.008 |
| Social Sciences | Predictive | 7.68 | $< 0.0001$ | 6.02 | $< 0.0001$ | 6.26 | 0.044 | 9.39 | 0.009 |
| $n = 21$ | Predictive and Heterogeneity | 4.30 | $< 0.0001$ | 4.03 | $< 0.0001$ | 3.81 | 0.15 | 5.11 | 0.078 |
| | Shrinkage | 6.00 | $< 0.0001$ | 4.86 | $< 0.0001$ | 4.49 | 0.11 | 6.14 | 0.046 |
| | Shrinkage and Heterogeneity | 2.79 | 0.005 | 2.80 | 0.005 | 2.88 | 0.24 | 3.33 | 0.19 |

A total of five miscalibration tests have been performed, a theoretically well-founded way to summarize the $p$-values of these tests is to use their harmonic mean (Good, 1958; Held, 2019b). For this reason, Table 3.5 shows the harmonic mean of the $p$-values from the score based tests and the Kolmogorov-Smirnov test of the PITs. Taken together, there is strong evidence for

**Table 3.6:** Results from all miscalibration tests, where the $p$-values haven been combined by the harmonic mean.

| Project | Method | $HM(p\text{-values})$ |
|---|---|---|
| Experimental Economics | Predictive | 0.011 |
| $n = 18$ | Predictive and Heterogeneity | 0.029 |
| | Shrinkage | 0.057 |
| | Shrinkage and Heterogeneity | 0.27 |
| Experimental Philosophy | Predictive | 0.0006 |
| $n = 31$ | Predictive and Heterogeneity | 0.51 |
| | Shrinkage | 0.0002 |
| | Shrinkage and Heterogeneity | 0.70 |
| Psychology | Predictive | $< 0.0001$ |
| $n = 73$ | Predictive and Heterogeneity | $< 0.0001$ |
| | Shrinkage | $< 0.0001$ |
| | Shrinkage and Heterogeneity | $< 0.0001$ |
| Social Sciences | Predictive | $< 0.0001$ |
| $n = 21$ | Predictive and Heterogeneity | $< 0.0001$ |
| | Shrinkage | $< 0.0001$ |
| | Shrinkage and Heterogeneity | 0.003 |

miscalibration of all methods in the psychology and social sciences data sets. In the economics data set, on the other hand, there is no evidence for miscalibration of the shrinkage method taking into account heterogeneity and weak evidence for miscalibration of the other methods. Finally, in the philosophy data set there is moderate evidence for miscalibration of the methods not taking into account heterogeneity and no evidence for miscalibration of the methods taking into account heterogeneity.

### 3.2.2 Binary predictions

Since statistical significance of the replication study is one of the most commonly used criteria for replication success (Klein *et al.*, 2014; Open Science Collaboration, 2015; Camerer *et al.*, 2016; Ebersole *et al.*, 2016; Camerer *et al.*, 2018; Cova *et al.*, 2018; Klein *et al.*, 2018), in the following section the probability of significance under the investigated predictive distributions will be evaluated using methods suited for binary predictive distributions. If not explicitly mentioned, the significance threshold $\alpha = 0.05$ for a two-sided $p$-value will be used.

#### Predicted probability of statistical significance

Figure 3.8 shows the probabilities of a statistically significant test statistic in the replication study under the investigated predictive distributions, grouped by whether or not the replications actually achieved significance. When looking at the statistical methods, the predicted probabilities of significance are generally high, even for many of the studies where the replications did not achieve significance. Comparing the different replication projects, in the social science data set the distributions of the predicted probabilities among all methods are virtually identical between the significant and non-significant replication studies, suggesting low discriminatory power of all methods. In the economics, philosophy, and psychology data sets, on the other

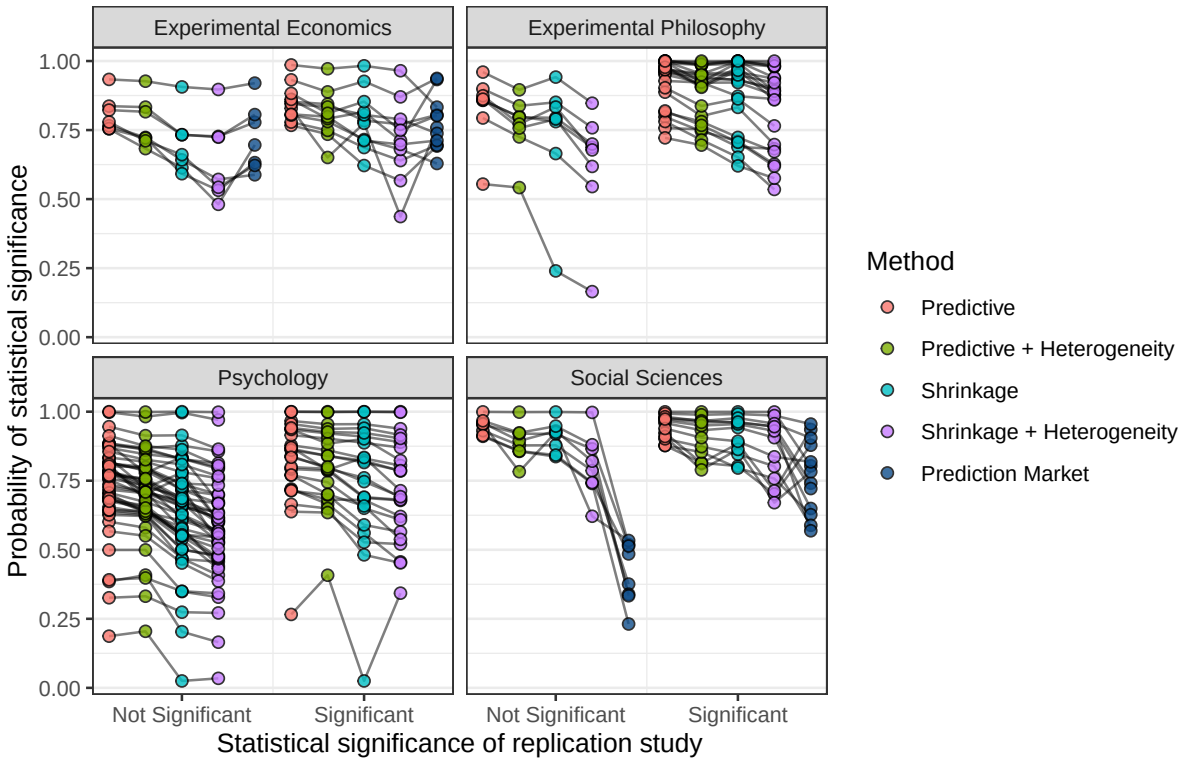


**Figure 3.8:** Probabilities of statistically significant replication outcome under predictive distributions (at $\alpha = 0.05$).

hand, the predicted probabilities of the non-significant replications are in most cases slightly smaller, indicating some discriminatory power of the predictions. Looking at the different prediction methods, the predicted probabilities are generally smaller for the shrinkage compared to the predictive methods, and similarly for methods taking into account heterogeneity compared to methods not taking into account heterogeneity. Moreover, in the social sciences data set the probabilities from the non-statistical prediction market method are much lower for non-significant replications compared to the probabilities of the significant replications, suggesting substantial discriminatory power of this method. In the economics data set, however, the prediction market probabilities are high for both significant and non-significant replications, indicating only low discriminatory power.

Figure 3.9 shows plots of the proportion of actually significant replications within the quartiles of the empirical distribution of predicted probabilities for significance. A loess smoother is underlaid to facilitate visual comparison. Ideally, the empirical proportions would monotonically increase from zero to one, however, in the economics, social sciences, and philosophy data sets, for certain methods there are lower quartiles with larger proportions of significant replications than higher quartiles. In the psychology data set, where the sample size is also much bigger than in the other data sets, the proportions look much more stable. Nevertheless, the proportion of significant replications is quite low even in the highest quartile, suggesting miscalibration.

### Expected number of statistically significant replication studies

By summing up all probabilities within one data set and method, the expected number of statistically significant replication outcomes is obtained and can be compared to the observed number, *e. g.* by using a $\chi^2$-goodness-of-fit test, as is shown in Table 3.7. In general, the observed number of significant replication studies is smaller than the expected number for all methods in all data sets, yet the amount of overestimation differs between the methods. Looking at the statisti-

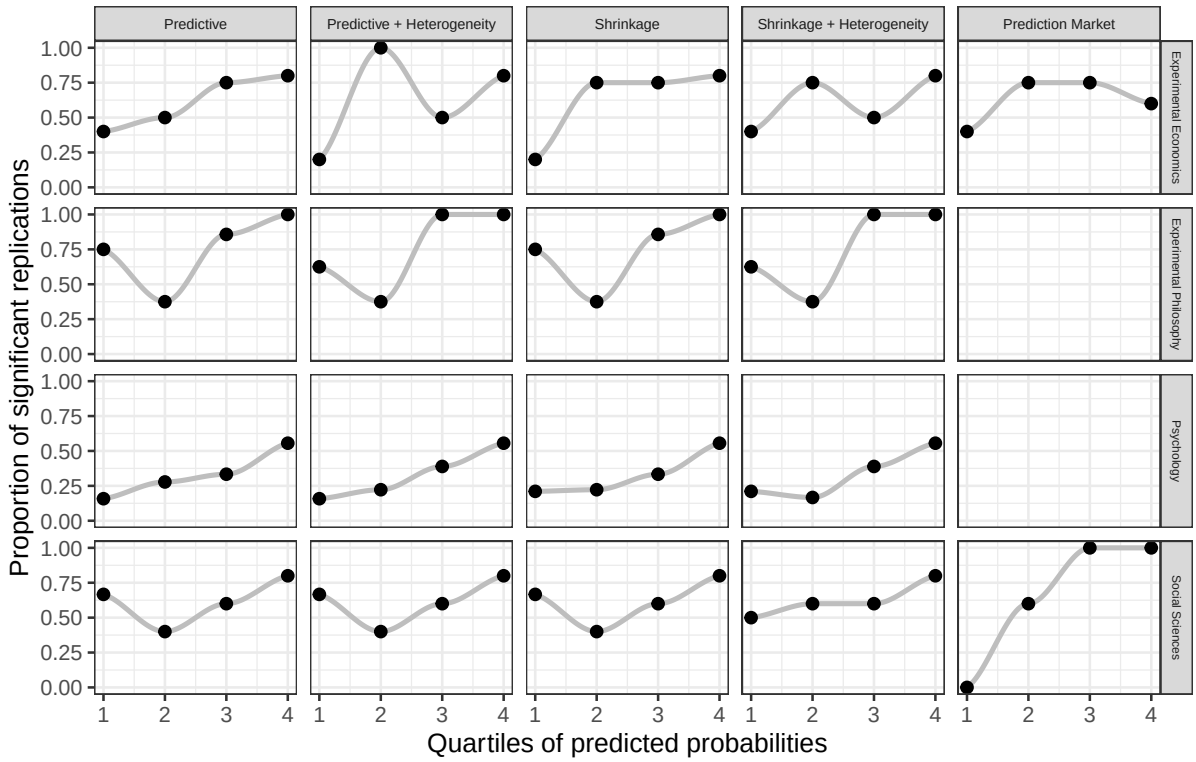


**Figure 3.9:** Calibration plot for binary predictions where $\alpha = 0.05$.

**Table 3.7:** Expected and observed number of statistically significant replication studies (at $\alpha = 0.05$).

| Project | Method | Observed | Expected | $p$-value |
|---|---|---|---|---|
| Experimental Economics | Predictive | 11 | 15.0 | 0.012 |
| $n = 18$ | Predictive and Heterogeneity | 11 | 14.3 | 0.057 |
| | Shrinkage | 11 | 13.6 | 0.16 |
| | Shrinkage and Heterogeneity | 11 | 12.4 | 0.49 |
| | Prediction Market | 11 | 13.6 | 0.16 |
| Experimental Philosophy | Predictive | 23 | 27.8 | 0.004 |
| $n = 31$ | Predictive and Heterogeneity | 23 | 26.5 | 0.076 |
| | Shrinkage | 23 | 26.2 | 0.11 |
| | Shrinkage and Heterogeneity | 23 | 24.1 | 0.65 |
| Psychology | Predictive | 24 | 55.4 | $< 0.0001$ |
| $n = 73$ | Predictive and Heterogeneity | 24 | 53.5 | $< 0.0001$ |
| | Shrinkage | 24 | 49.2 | $< 0.0001$ |
| | Shrinkage and Heterogeneity | 24 | 45.9 | $< 0.0001$ |
| Social Sciences | Predictive | 13 | 19.9 | $< 0.0001$ |
| $n = 21$ | Predictive and Heterogeneity | 13 | 18.9 | $< 0.0001$ |
| | Shrinkage | 13 | 19.2 | $< 0.0001$ |
| | Shrinkage and Heterogeneity | 13 | 17.6 | 0.006 |
| | Prediction Market | 13 | 13.3 | 0.89 |

cal prediction methods, the overestimation is the smallest for the shrinkage method taking into account heterogeneity and the largest for the predictive method across all data sets. When comparing the different data sets, in the economics and philosophy data sets there is no evidence for

a difference between expected and observed under the shrinkage methods, whereas there is weak to moderate evidence of a difference for the predictive methods, suggesting better calibration of the former compared to the latter. In the social sciences and psychology data sets, on the other hand, there is strong evidence for a difference between the expected from the observed number of significant replications for all methods. Furthermore, the expected number under the prediction market method in the economics and social sciences data sets do not differ substantially from what was actually observed.

Propositions have been made recently for lowering the significance threshold for the claim of new scientific discoveries (Benjamin *et al.*, 2017). For this reason, it is also interesting to compare the expected and observed number of statistically significant replication outcomes for smaller significance thresholds than 0.05, as shown in Figure 3.10. For all values of $\alpha$, the expected

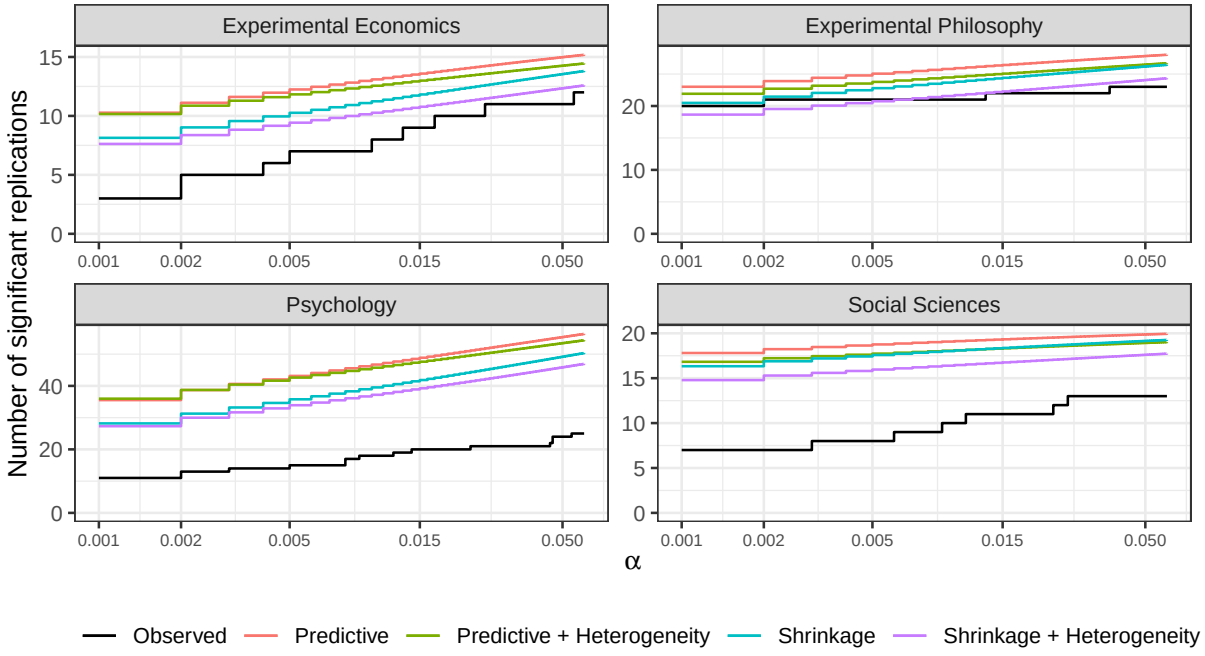


**Figure 3.10:** Expected and observed number of statistically significant replication studies as function of $\alpha$.

number is smaller for the shrinkage methods than for the predictive methods, and it is also smaller when taking into account heterogeneity compared to when not taking heterogeneity into account. These results indicate again that the shrinkage method taking into account heterogeneity leads to more realistic predictions. Comparing the different data sets, in the psychology and social sciences data sets the difference between the expected and observed number of significant replications is large across the whole range of possible significance thresholds for all four prediction methods. In the philosophy and the economics data set, on the other hand, the expected number is much closer to the observed number, especially for the shrinkage methods.

### Brier scores

In Table 3.8 the mean (normalized) Brier scores are shown for each data set and prediction method. In addition, in Figure 3.11 the Brier scores are shown visually with the corresponding standard errors. Note that the standard errors are only shown to illustrate the spread of the individual scores and not to compare the mean scores between the methods. Comparing the different data sets, in the social sciences and psychology data sets the binary predictive performance is poor for all statistical methods. Namely, the mean Brier scores of all statistical

**Table 3.8:** Mean (normalized) Brier scores of binary predictions, where $\alpha = 0.05$.

| | | Type | |
|---|---|---|---|
| Project | Method | BS mean | normalized BS mean |
| Experimental Economics | Predictive | 0.27 | −0.14 |
| $n = 18$ | Predictive and Heterogeneity | 0.26 | −0.10 |
| | Shrinkage | 0.23 | 0.05 |
| | Shrinkage and Heterogeneity | 0.23 | 0.05 |
| | Prediction Market | 0.24 | −0.02 |
| Experimental Philosophy | Predictive | 0.19 | −0.01 |
| $n = 31$ | Predictive and Heterogeneity | 0.17 | 0.11 |
| | Shrinkage | 0.17 | 0.10 |
| | Shrinkage and Heterogeneity | 0.15 | 0.23 |
| Psychology | Predictive | 0.39 | −0.78 |
| $n = 73$ | Predictive and Heterogeneity | 0.36 | −0.64 |
| | Shrinkage | 0.34 | −0.52 |
| | Shrinkage and Heterogeneity | 0.29 | −0.31 |
| Social Sciences | Predictive | 0.35 | −0.47 |
| $n = 21$ | Predictive and Heterogeneity | 0.31 | −0.32 |
| | Shrinkage | 0.32 | −0.37 |
| | Shrinkage and Heterogeneity | 0.27 | −0.16 |
| | Prediction Market | 0.11 | 0.52 |

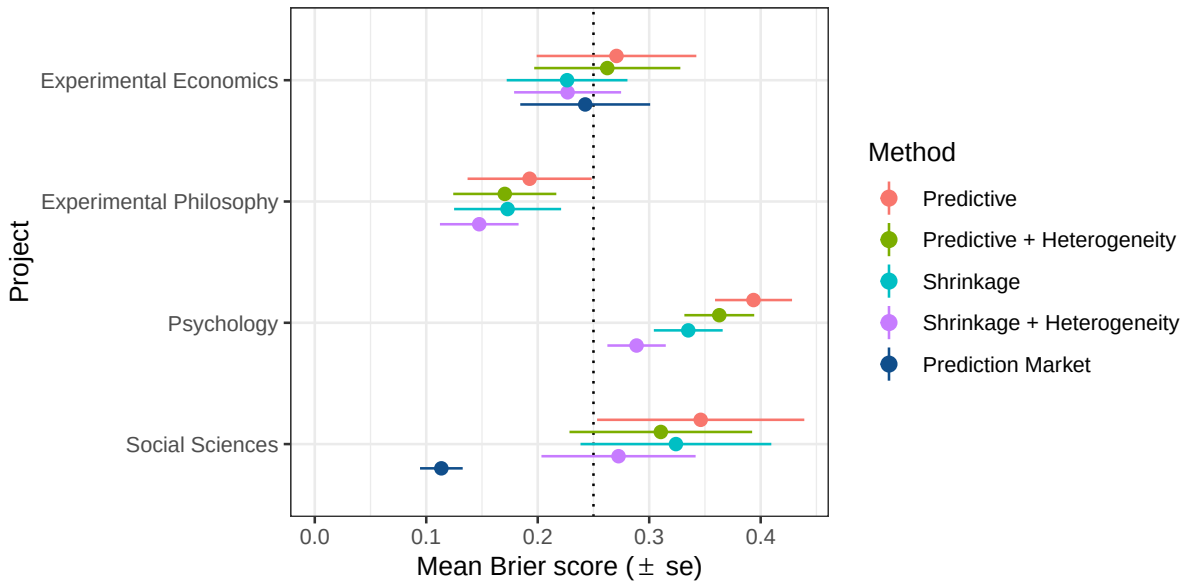


**Figure 3.11:** Mean Brier scores.

methods are larger than 0.25, a score that can be obtained by simply using 0.5 every time as prediction probability. Looking at the mean normalized Brier score, performance is even worse because a useful prediction must be at least as good as using the prevalence of the binary event as its prediction probability. It should be noted, however, that this comparison is in some sense unfair, since a prediction is based only on one original study and not on an entire sample of studies. In the economics data set, on the other hand, both shrinkage methods achieve a

positive mean normalized Brier score, while it is negative for the predictive methods. Finally, the predictions in the philosophy data set show the best performance, *i. e.* all methods except the predictive method achieve a positive mean normalized Brier score with the shrinkage method taking into account heterogeneity showing the largest value, suggesting that this method provides the best predictions. Moreover, in the economics data set the prediction market method shows a normalized score of about zero, which is comparable with the statistical prediction methods. The prediction market predictions in the social sciences data set, however, show an extremely good performance, far better than all statistical methods in this data set.

It is also possible to compute the mean Brier scores for binary predictions at other significance thresholds $\alpha$ than 0.05, as shown in Figure 3.12. Across the whole range of $\alpha$ values, the shrinkage

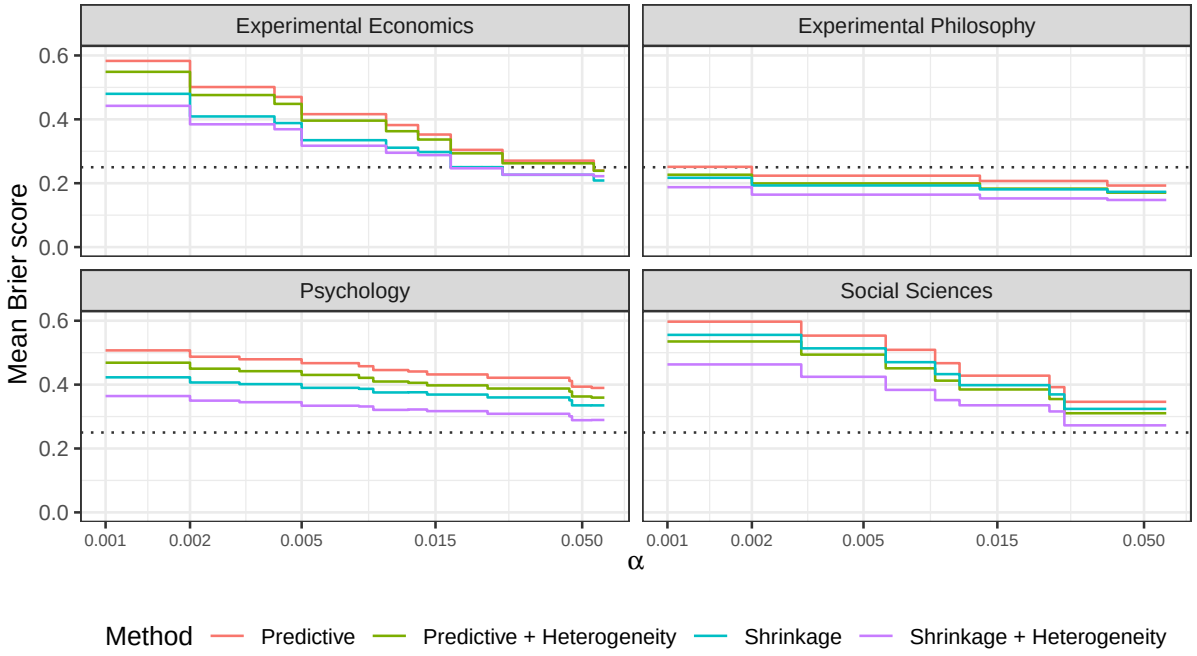


**Figure 3.12:** Mean Brier score as function of $\alpha$.

method taking into account heterogeneity shows the smallest mean Brier scores in all data sets, indicating the superiority of this method. However, the mean Brier scores increase with lower values of $\alpha$ for all methods, suggesting that they overestimate the probability of significance also for lower thresholds. Only in the philosophy data set, the mean Brier scores remain below 0.25 for all but the predictive method. In the other data sets the mean Brier scores exceed or never reach values below 0.25.

Table 3.9 shows the results of Spiegelhalter's $z$-test. In the psychology and social sciences data sets the tests provide strong evidence for miscalibration of all statistical prediction methods, but no evidence for miscalibration of the prediction market method in the social sciences data set. In the economics data set, on the other hand, there is no evidence for miscalibration of both shrinkage and the prediction market method and weak evidence for miscalibration of the predictive methods. Finally, in the philosophy data set there is moderate evidence for miscalibration of the predictive method and weak evidence for miscalibration of the shrinkage method not taking into account heterogeneity, but no evidence for miscalibration of both methods taking into account heterogeneity.

**Table 3.9:** Results from Brier score based miscalibration tests (Spiegehalter's $z$-test).

| Project | Method | $z$ | $p$-value |
|---|---|---|---|
| Experimental Economics | Predictive | 2.49 | 0.013 |
| $n = 18$ | Predictive and Heterogeneity | 2.03 | 0.042 |
| | Shrinkage | 1.15 | 0.25 |
| | Shrinkage and Heterogeneity | 0.79 | 0.43 |
| | Prediction Market | 0.70 | 0.48 |
| Experimental Philosophy | Predictive | 3.31 | 0.0009 |
| $n = 31$ | Predictive and Heterogeneity | 1.60 | 0.11 |
| | Shrinkage | 2.07 | 0.039 |
| | Shrinkage and Heterogeneity | 0.21 | 0.83 |
| Psychology | Predictive | 10.00 | $< 0.0001$ |
| $n = 73$ | Predictive and Heterogeneity | 8.50 | $< 0.0001$ |
| | Shrinkage | 7.81 | $< 0.0001$ |
| | Shrinkage and Heterogeneity | 5.17 | $< 0.0001$ |
| Social Sciences | Predictive | 7.20 | $< 0.0001$ |
| $n = 21$ | Predictive and Heterogeneity | 4.89 | $< 0.0001$ |
| | Shrinkage | 5.46 | $< 0.0001$ |
| | Shrinkage and Heterogeneity | 3.42 | 0.0006 |
| | Prediction Market | $-1.14$ | 0.25 |

### Calibration slope

Figure 3.13 shows the calibration slopes obtained by logistic regression of the outcome whether the replication achieved statistical significance on the logit transformed predicted probabilities. In all data sets except the psychology one, the confidence intervals are very wide due to the small

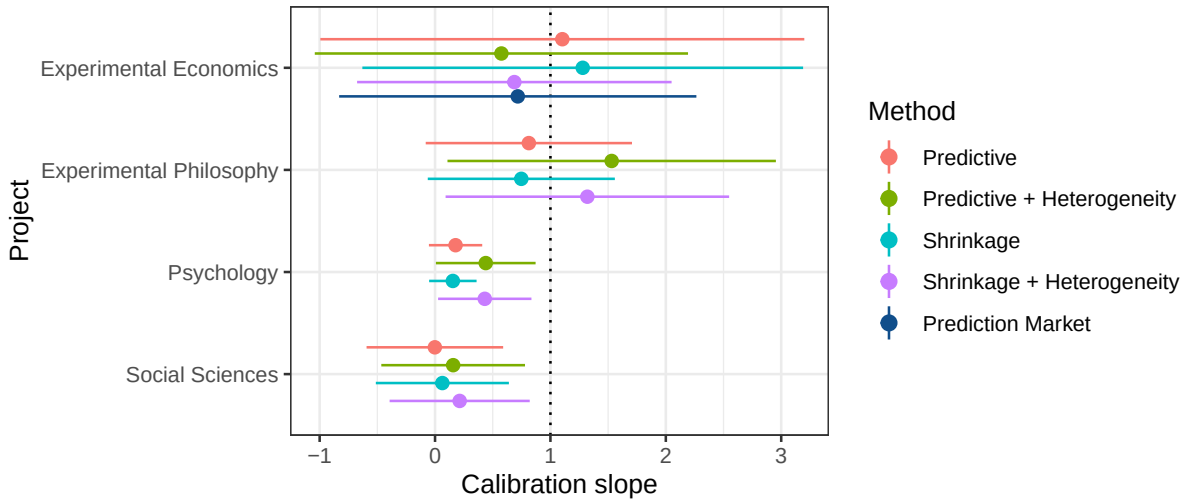


**Figure 3.13:** Calibration slope for binary predictions with 95% confidence interval (at $\alpha = 0.05$).

sample size. Also note that the prediction market method in the social science data set had numerical problems because of perfect separation, which is why the corresponding calibration slope is not shown. Looking at the psychology and social sciences data sets, the calibration slopes of all methods are considerably below the nominal value of one, suggesting that the predicted probabilities are too high. However, the shrinkage and predictive methods which take

into account heterogeneity show values closer to one than the methods that do not take into account heterogeneity, suggesting better calibration of the former compared to the latter. In the economics data set the methods not taking into account heterogeneity show slightly larger values than one, while the methods which do not take into account heterogeneity as well as the prediction market method show smaller values than one. On the other hand, in the philosophy data set the reverse is visible, *i. e.* heterogeneity methods show larger values than one, whereas the methods not taking into account heterogeneity shows smaller values than one.

For the statistical prediction methods it is also possible to compute calibration slopes for other significance thresholds $\alpha$, as shown in Figure 3.14. Looking at the psychology and social sciences

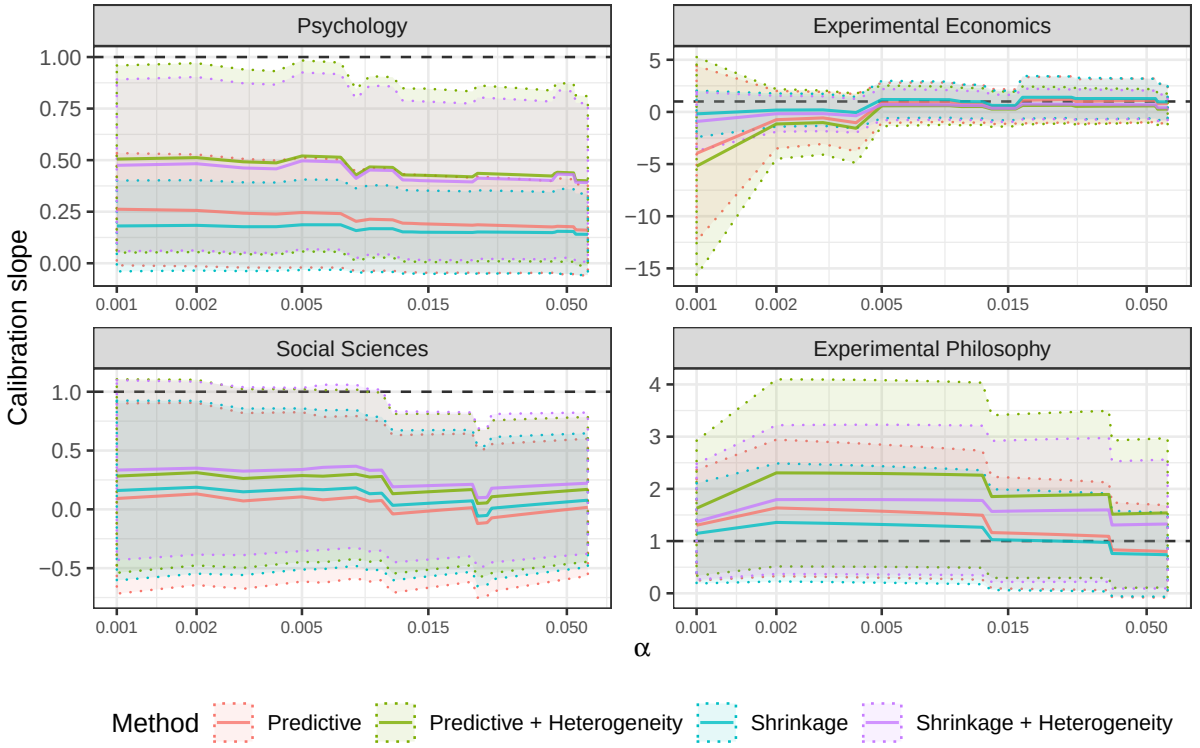


**Figure 3.14:** Calibration slopes with 95% confidence intervals as function of $\alpha$.

data sets, all calibration slopes increase slightly for lower values of $\alpha$, with the heterogeneity methods showing the highest values, yet all still remain below the nominal value of one. In the the economics data set, one the other hand, the calibration slopes become smaller with decreasing $\alpha$ for all methods, the predictive methods show even negative values. Furthermore, for decreasing values of $\alpha$ the confidence intervals of the calibration slope become extremely wide. Finally, in the philosophy data set the slopes of all methods increase with decreasing $\alpha$ until $\alpha = 0.002$, where they start to decrease again to values close to one.

### Area under the curve

Figure 3.15 shows the area under the curve (AUC) of the binary predictions under the respective predictive distributions. The 95% Wald type confidence intervals were computed on the logit scale and then backtransformed. Note that in the social sciences data set for the prediction market method, an AUC of one was obtained because the predictions were able to perfectly separate non-significant and significant replications. For this reason, it was also not possible to compute a confidence interval with the method used. The statistical prediction methods in the social sciences data set, on the other hand, show AUCs between 0.5 and 0.6 with wide confidence intervals, suggesting no discriminatory power. In the philosophy and psychology data sets the

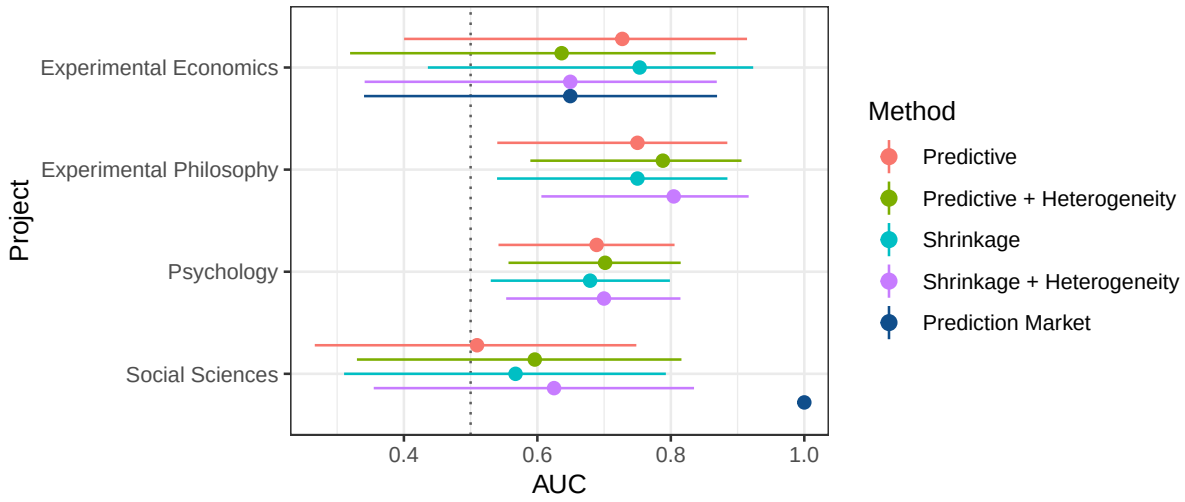


**Figure 3.15:** Area under the curve (AUC) with 95% confidence interval (at $\alpha = 0.05$).

methods that take into account heterogeneity show the highest AUCs. The AUCs of the former are around 0.8, while they are about 0.7 for the latter, indicating some discriminatory power of the predictions in both data sets. Finally, in the economics data set the non-heterogeneity methods achieve the highest AUCs with values of around 0.75, but with very wide confidence intervals.

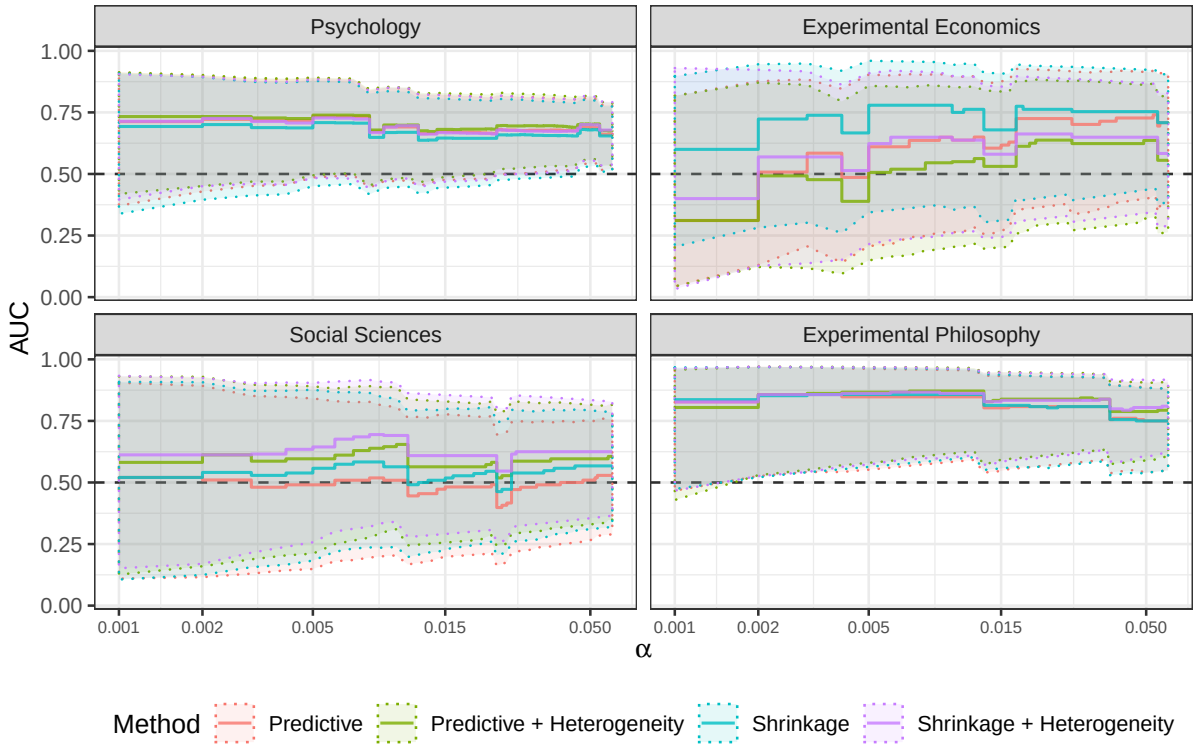


**Figure 3.16:** Area under the curve with 95% confidence intervals as function of $\alpha$.

Figure 3.16 shows the AUC as a function of the significance threshold $\alpha$. Due to fewer replications that are significant for small $\alpha$, the confidence intervals become wider as $\alpha$ decreases. In the philosophy, social sciences, and psychology data sets, the AUCs of all methods stay more or less constant over the entire range of $\alpha$. Namely, the AUCs of all methods remain around 0.5

to 0.6 in the social sciences data set, while they remain around 0.7 to 0.8 in the psychology and philosophy data sets. In the economics data set, on the other hand, for all methods the AUCs also decrease with decreasing $\alpha$ to values of around 0.5.

## 3.3 Sensitivity analysis of heterogeneity parameter choice

For the methods which take into account between study heterogeneity, the heterogeneity parameter $\tau$ was set to a value of 0.08 through theoretical considerations. For this reason, it is advisable to perform a sensitivity analysis to investigate how much the results change when other values for $\tau$ are selected. In this section, the change in predictive performance for other values of $\tau$ will be primarily investigated using the mean score, since it provides a good summary measure for calibration and sharpness of a predictive distribution (Gneiting and Katzfuss, 2014).

Assuming that the between study heterogeneity for studies within one field is similar, one can look at the mean score as a function of $\tau$ when setting $\tau$ to the same value for all studies within a project, as shown in Figure 3.17. In each plot, the minimum values are indicated by a cross and are also displayed numerically. In general, many of the mean score functions look

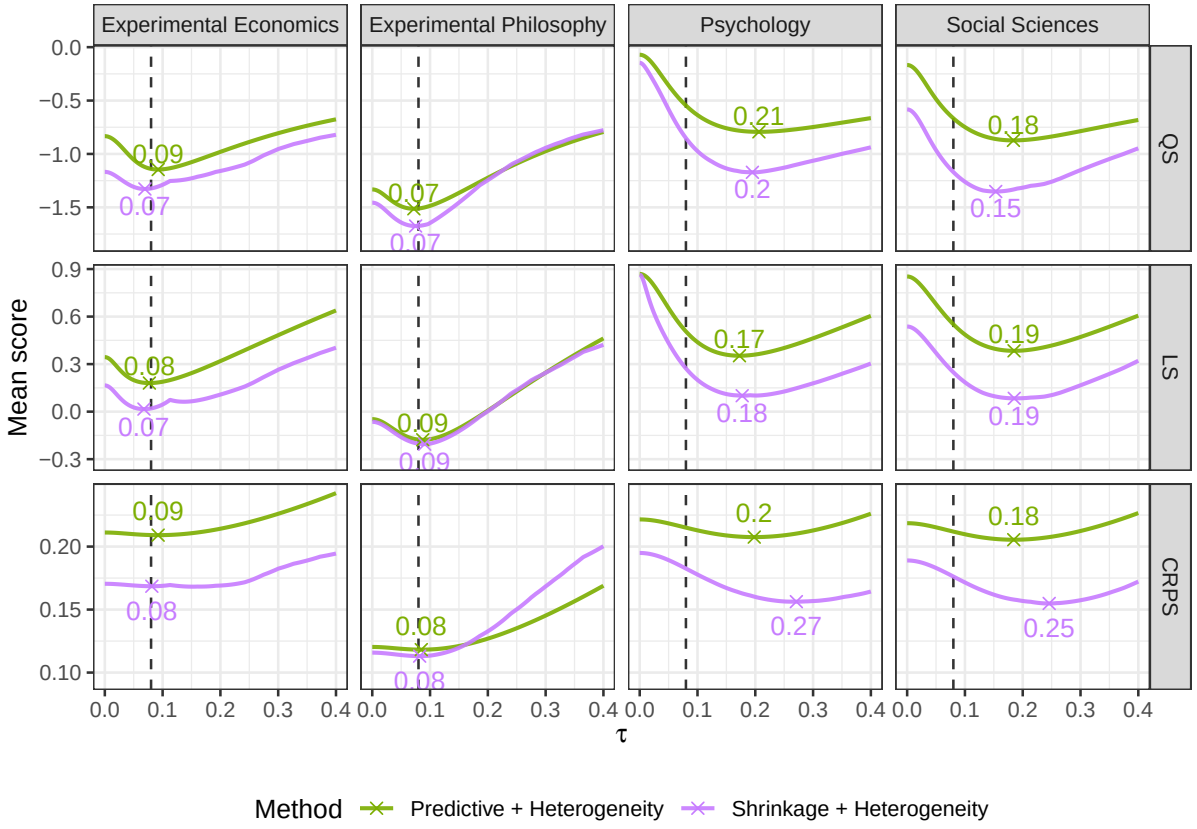


**Figure 3.17:** Mean scores as a function of $\tau$ for each score type and project. The dashed line indicates the chosen value of 0.08.

rather flat, suggesting large uncertainty about the $\tau$ parameter. Comparing the two models, for the same value of $\tau$, the shrinkage model shows smaller mean scores over the entire range of $\tau$ than the predictive model in all but the philosophy data set. In the philosophy data set, on the other hand, both models usually show comparable mean scores. This suggests that shrinkage leads to a better (or at least equal) predictive performance across all data sets, regardless of the choice of $\tau$. Looking at the different score types, the shape of the quadratic and logarithmic mean score functions looks very similar within each data set. The mean CRP score functions,

however, usually look flatter and their minima differ in some cases from the other two score types. Namely, in the social sciences and psychology data sets the $\tau$ values which minimize the mean CRP score functions of the shrinkage model are substantially larger compared to the predictive model as well as to the other score types. When comparing the minima across data sets, in the economics and philosophy data sets all minima are slightly below 0.1, which is close to the theoretically motivated value of 0.08 that was chosen. In the psychology and social sciences data sets, on the other hand, most of the minima are slightly below 0.2, which is much higher than the chosen value of 0.08.

Figure 3.18 shows the PIT histograms of the predictions using the $\hat{\tau}$ that minimize the mean logarithmic score for each project. The logarithmic score was selected because it is a local

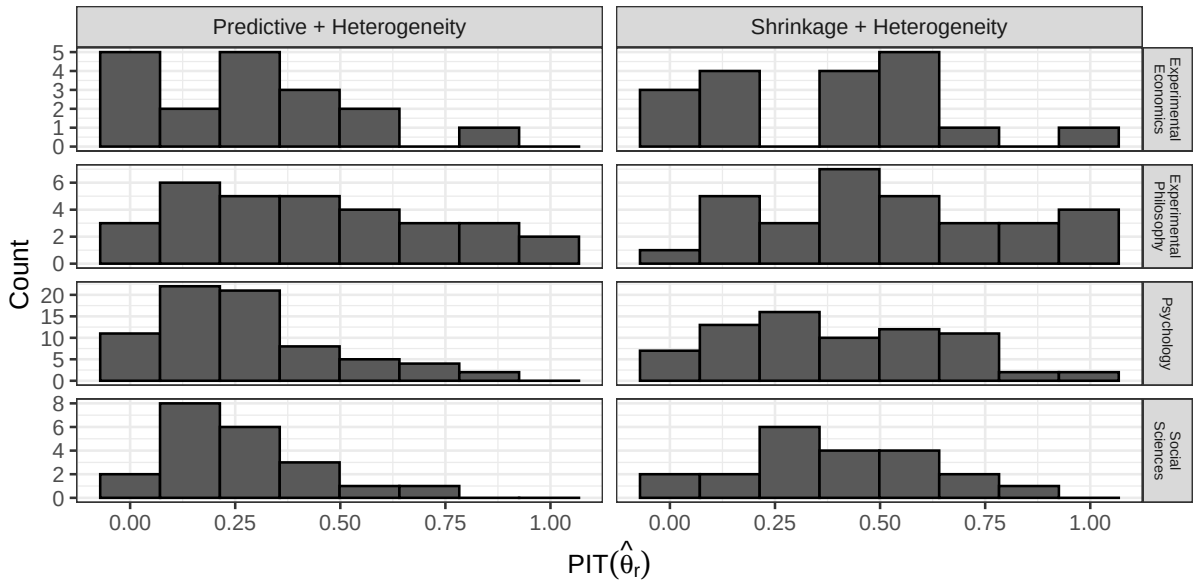


**Figure 3.18:** PIT histograms of predictions using $\hat{\tau}$ that minimize the mean score per project.

scoring rule, which is the preferred type of scoring rule for inference of a parameter (see p. 72 in Bernardo and Smith, 2000). Compared to the PIT histograms in Figure 3.6, the distribution of the PIT values did not change much for the economics and philosophy data sets, since the $\hat{\tau}$ which minimize the logarithmic mean score are almost identical to the initially chosen value of 0.08. However, the histograms of the predictions in the psychology and social sciences data sets look flatter, especially the ones where the shrinkage method was used, suggesting improved calibration compared to the predictions using the initially chosen values of $\tau$.

Miscalibration of the predictions using the $\hat{\tau}$ which minimize the logarithmic mean score was also investigated using miscalibration tests. Table 3.10 shows the harmonic mean of the $p$-values from the four scoring rule based miscalibration test (unconditional LS and CRPS tests, regression based DSS and CRPS tests) and the Kolmogorov-Smirnov test of the PITs. There is only evidence for miscalibration of the predictive methods in the psychology and social sciences data sets and weak evidence for miscalibration of the shrinkage method in the psychology data set and the predictive method in the economics data set. Thus, by increasing the value of $\tau$ for the predictions in the social sciences and psychology data sets, miscalibration could be drastically reduced, especially for the shrinkage predictions. Nevertheless, it should be noted that these results are purely exploratory and should be viewed with caution since the same data have been used twice.

One can go one step further and estimate $\tau$ for each study pair individually, as shown in Figure 3.19. However, the uncertainty with these estimates is large, because for each estimate

**Table 3.10:** Harmonic mean of $p$-values from miscalibration tests of predictions using $\hat{\tau}$ that minimize the mean score per project.

| Project | Method | $HM(p\text{-values})$ |
|---|---|---|
| Experimental Economics | Predictive and Heterogeneity | 0.03 |
| $n = 18$ | Shrinkage and Heterogeneity | 0.24 |
| Experimental Philosophy | Predictive and Heterogeneity | 0.54 |
| $n = 31$ | Shrinkage and Heterogeneity | 0.84 |
| Psychology | Predictive and Heterogeneity | $< 0.0001$ |
| $n = 73$ | Shrinkage and Heterogeneity | 0.027 |
| Social Sciences | Predictive and Heterogeneity | 0.0003 |
| $n = 21$ | Shrinkage and Heterogeneity | 0.23 |

there is only one study pair. Moreover, in the case of the shrinkage predictive distribution, numerical optimization is not guaranteed to identify the global minimizer $\hat{\tau}$ since the objective is generally not convex (the score functions would have to be visually assessed for each study pair and each score type to confirm the identification of the global minimum). It is nevertheless interesting to compare the study pair specific estimates to their project wise counterparts. In

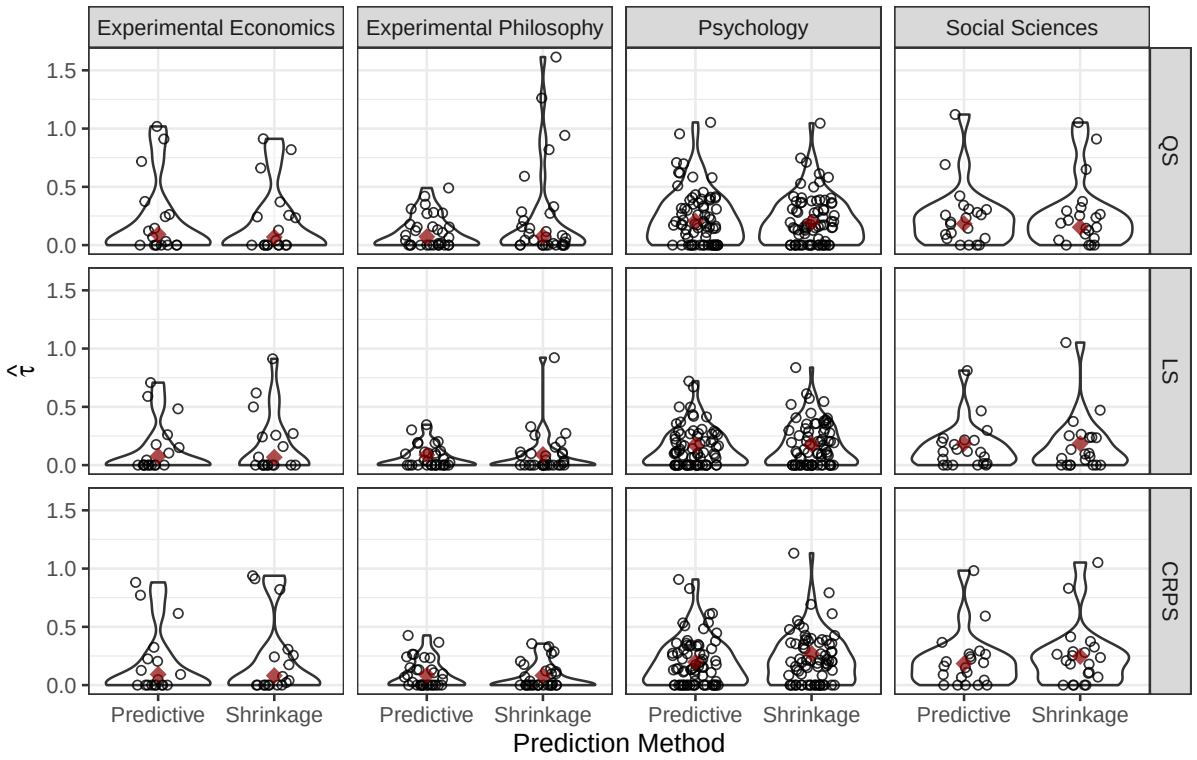


**Figure 3.19:** Estimates $\hat{\tau}$ which minimize score for each study pair and score type. The value of $\tau$ which minimizes the mean score project-wise is depicted by a red diamond.

most projects, there are some study pairs whose estimates are exactly zero or extremely large, but most of the estimates are scattered around the values which minimize the project-wise mean scores. Using these estimates to compute predictions, of course, the mean scores become even lower compared to the project-wise estimates. However, reasonable improvements in calibration could already be achieved using the project-wise estimates, which is why further evaluations of the study pair specific estimates are not shown.

# Chapter 4

# Discussion

This thesis addressed the question to what extent it is possible to predict the effect estimate of a replication study using the effect estimate from the original study and knowledge of the sample size in both studies. For this purpose, several models of effect sizes were discussed and adapted to the setting of replication studies. In all models it was assumed that after a suitable transformation an effect can be modelled by a normally distributed random variable. Furthermore, the models either assumed that the original study had identified the effect correctly (*predictive model*), or they shrunk the effect towards zero based on the evidence in the original study (*shrinkage model*). In a Bayesian framework, the former can be achieved by a flat prior distribution of the effect, while the latter is achieved by using a zero-mean normal prior and estimating the variance parameter by empirical Bayes. Finally, the models also differed in terms of whether between study heterogeneity of the effects was taken into account or not, which was achieved by a hierarchical model structure of the effect sizes.

Patil *et al.* (2016) examined a similar research question and computed predictions of replication effect estimates using the data set from the replication project psychology (Open Science Collaboration, 2015). The model used by Patil *et al.* (2016) was derived in a non-Bayesian framework, but corresponds to the most basic model derived in this thesis (*i. e.* the effect identified in the original study is correct, no between study heterogeneity). The same method was also used in the analyses of the experimental economics replication project (Camerer *et al.*, 2016) and the social sciences replication project (Camerer *et al.*, 2018). In all of these analyses, however, apart from examining the coverage of the prediction intervals, no systematic evaluation of the predictive distributions was conducted, even though there exist many well established methods for evaluating probabilistic predictions (for an overview see *e. g.* Gneiting and Katzfuss, 2014).

Therefore, using the four different prediction models, the corresponding predictive distributions were calculated for the three aforementioned data sets and for the data from the experimental philosophy replicability project (Cova *et al.*, 2018). Calibration, sharpness, and discrimination of these predictions were then evaluated using established methodology, such as proper scoring rules, probability integral transform, calibration slope, and area under the curve.

## 4.1 Predictive evaluation

By taking into account between study heterogeneity and/or evidence based shrinkage, calibration and sharpness could be improved compared to the predictive method. That is, the predictions obtained with the shrinkage and heterogeneity method usually showed a higher coverage of the prediction intervals, more uniformly distributed PIT values, substantially lower mean scores, and less or no evidence of miscalibration. The improvements in predictive performance by using the shrinkage and heterogeneity method were usually the largest in the social sciences and psychology and the smallest in the economics and philosophy data sets. However, in the psychology and social sciences data sets, the tests still suggest some miscalibration, even for the heterogeneity

and shrinkage model which performed the best, while there is less evidence for miscalibration in the philosophy and economics data sets.

Since statistical significance of the replication study is a commonly used criterion for replication success, the probability of significance under the investigated prediction models was specifically evaluated. In general, predictive performance under the shrinkage and heterogeneity model improved compared to the predictive model also when the objective was to predict significance. Namely, in the economics and philosophy data set, the expected number of significant replication studies was reasonably close to the observed number, the mean Brier scores were lower, less or no miscalibration could be detected, and area under the curve and calibration slope indicate some predictive power. Similarly, for the psychology and social sciences data sets, the evaluation methods mostly indicate improvements in discrimination and sharpness of the shrinkage and heterogeneity models compared to the predictive model. Despite these improvements, however, the methods still showed miscalibration and the predictive performance was generally worse compared to the economics and philosophy data sets. Furthermore, in the social sciences and economics data sets, the predictions could be compared to the non-statistical prediction market method which provides an estimate of the peer-beliefs about the probability of significance. In the economics data set, the shrinkage methods showed equal performance compared to the prediction market, while in the social sciences data set, the prediction market method performed better than any of the statistical methods. Finally, the evaluation of the binary predictions was also conducted for smaller significance thresholds than the conventional 0.05 level. The results do not indicate better predictive performance for lower thresholds, but the evaluation methods often become unstable, which may be due to the smaller number of studies that are significant at lower thresholds.

To conclude, it seems likely that in many of the investigated data there is between study heterogeneity present, since the models that take heterogeneity into account always showed equal or better performance compared to their non-heterogeneity counterparts. This is not surprising, as many of the replications used samples from different populations and/or different materials than those in the original studies. Evidence based shrinkage also improved predictive performance in most cases, suggesting that many of the effect estimates from the original studies were either inflated or false positives. Possible reasons for this might be publication bias or the use of questionable research practices.

## 4.2 Sensitivity analysis of heterogeneity parameter choice

It is not possible to estimate the heterogeneity parameter $\tau$ using only data from one original study. For this reason, a sensible value of $\tau$ had to be determined on the basis of theoretical arguments. Namely, $\tau$ was specified such that the difference of the 97.5% to the 2.5% quantile of a correlation effect size with mean zero is approximately the size of a medium effect size according to the Cohen (1992) classification. This classification is established in practice, but also sometimes criticized, *e. g.* recently in Funder and Ozer (2019).

A sensitivity analysis was conducted to examine the impact of this decision on the results. The relatively flat mean score functions indicate that there is a lot of uncertainty about the between study heterogeneity parameter. However, the chosen value of 0.08 seems plausible for the economics and philosophy data sets, as it was close to the minima of all mean score functions. On the other hand, the values of $\tau$, which minimized the mean score functions in the social sciences and psychology data sets were substantially larger than 0.08. Nevertheless, for the same value of $\tau$, the mean scores under the shrinkage method were smaller or equal compared to the predictive method across most score types and data sets. This suggests that the comparison of the methods is not severely influenced by the choice of $\tau$. Furthermore, when choosing higher values of $\tau$ for the predictions in the social sciences and psychology data sets, calibration could still be improved, illustrating the high flexibility of the shrinkage and heterogeneity model.

## 4.3 Differences between replication projects

The predictive performance of the investigated methods varied between the data sets. Usually, the performance was the best in the philosophy and economics data sets and the worst in the psychology and social sciences data sets.

There are several possible explanations for this phenomenon. The number of studies within a replication project could be one possible reason for the differences in the results of some of the evaluation methods, *e. g.* miscalibration tests. That is, the psychology data set consists of many more study pairs, which leads to higher power to detect miscalibration in this data set compared to the other data sets. However, the social sciences data set only consists of a small number of studies and the tests still suggested miscalibration of all methods.

A further explanation might be that differences in the study selection process of the replication projects lead to the differences in predictive performance. For instance, the original studies in the philosophy, economics, and psychology projects were selected from ordinary journals from their respective fields. In the social sciences project, on the other hand, they were selected from the journals *Nature* and *Science*, which are known to mainly promote "innovative and exciting" research. Furthermore, if an original study contained several experiments, the rules to select the experiment to be replicated differed between the projects. In the psychology project, by default the last experiment was selected, whereas in the social sciences and philosophy projects by default the first experiment was selected. In the economics project, however, "the most central result" according to the judgement of the replicators was selected by default. If on average researchers report more robust findings at the first position and more exploratory findings at the last position of a publication (or the other way around), this might have systematically influenced the outcome of the replication studies. Similarly, when replicators can decide for themselves which experiment they want to replicate, they might systematically choose experiments with more robust effects that are easier to replicate.

Another possibility might be that the degree of inflation of the original study effect estimates varies between the different fields and leading to differences in overall predictive performance. In particular, in the economics, social sciences, and psychology data sets, the predictive performance was more substantially improved through evidence based shrinkage than in the philosophy data set. Cova *et al.* (2018) argue that the smaller inflation of effect estimates in experimental philosophy research may be caused by differences in academic culture between the fields. Namely, experimental philosophy is a much younger field where there is higher acceptance for negative or null results, which might make it less susceptible to publication bias than fields with more traditional "publish-or-perish" cultures, such as psychology.

## 4.4 Conclusions

The systematic evaluation of predictions of replication study outcomes provided new insights. Namely, by using a model of effect sizes which can take into account inflation of original study effect estimates and between study heterogeneity, it was possible to predict the effect estimate of the replication study with good predictive performance in two of the four data sets. In the other two data sets, predictive performance could still be drastically improved compared to the model proposed by Patil *et al.* (2016), which assumes that the effect estimate of the original study is not inflated and that there is no between study heterogeneity.

These results have various implications. First, for the assessment of probabilistic predictions of replication outcomes, data analysts should use more appropriate methods. For continuous predictions they should not only examine the coverage of prediction intervals as it was done in Patil *et al.* (2016); Camerer *et al.* (2016, 2018), but evaluate calibration and sharpness specifically. Similarly, for the evaluation of binary predictions, such as peer-beliefs estimated through prediction markets, methods to assess discrimination, calibration, and sharpness can provide more

valuable insights compared to the correlation tests used in Camerer *et al.* (2016, 2018). Second, the developed model could also be used to determine the sample size of a new replication study, considering potentially inflated and heterogeneous effect estimates, which seems realistic in view of the found results. This method would provide a more justified approach in comparison to just shrinking the target effect size ad hoc by an arbitrary amount as it was often done in the planning of previous replication studies (*e. g.* in Camerer *et al.*, 2018). Furthermore, in contrast to the classical power, the developed method also takes into account the uncertainty of the effect estimate from the original study. A ready to use R function is available in Appendix A.1. Finally, it is not a good idea to reduce replication success solely to whether a replication study achieves statistical significance or not. From a predictive point of view, it is often very likely that non-significance will occur, even if the underlying effect is not zero. Researchers should instead adopt more quantitative and probabilistic reasoning to assess replication success. Methods such as replication Bayes factors (Ly *et al.*, 2018) or the sceptical $p$-value (Held, 2019a) are promising approaches to replace statistical significance as the main criterion for replication success.

Moreover, these results offer interesting insights about the predictability of replication outcomes in four different fields. However, they should not be interpreted in such a way that research from one field is more credible than research from another. There are many other factors which could explain the observed differences in predictive performance, *e. g.* selection bias or small sample size. The complexity underlying any replication project is enormous, we should applaud to all the researchers involved for investing their limited resources in this endeavours. There is an urgent need to develop new methods for the design and analysis of replication studies, these data sets will be particularly useful for these purposes.

The approach used in this thesis also has some limitations. In all models, the simplifying assumption of normally distributed likelihood and prior has been made, which of course can be questionable for smaller sample sizes. It would be interesting to generalize the methods to other settings, for instance the $t$-distribution. Furthermore, the data sets used come all from relatively similar fields of academic science. It would also be of interest to perform the same analysis on data from "harder" scientific fields, such as physics, chemistry, or biology, as well as for non-academic research. Also the data from the three "Many Labs" projects (Klein *et al.*, 2014; Ebersole *et al.*, 2016; Klein *et al.*, 2018) were not considered, because the design of these projects differs drastically from the other four projects. Namely, in the "Many Labs" projects a smaller number of original studies were selected, the experiments of these studies were then combined into a single experiments, of which replications were conducted by multiple collaborators across the globe. It would be interesting to conduct the analyses also on these data, especially for the assessment of heterogeneity. Finally, the data from the prediction market of the reproducibility project psychology (Dreber *et al.*, 2015) were not evaluated. This was done because these data included only a subset of 44 studies that different from the meta-analytic subset used for this thesis and because they were not straightforward to combine with the other data about the effect estimates and sample sizes of the corresponding studies. It would also be of interest to compare these prediction market data with the statistical methods.

# Appendix A

# R code

## A.1 Prediction and evaluation methods

```r
# ============================================================================
# Compute sample size under predictive/shrinkage & heterogeneity model
# ============================================================================
sampleSizeReplication <- function(t_o, d = 0, power = 0.8, level = 0.05,
                                  alternative = "two.sided", prior = "flat") {

  ## Specify direction and critical value
  direction <- ifelse(t_o >= 0, 1, -1)
  alt <- ifelse(alternative == "two.sided", 2, 1)
  critical_value <- direction*qnorm(1 - level/alt)

  ## Define power function depending on prior
  if (prior == "flat") {
    s <- 1
  }
  if (prior == "sceptical") {
    s <- pmax(1 - (1 + d)/t_o^2, 0)
  }
  powerFun <- function(c) {
    power <- pnorm(q = critical_value,
                   mean = s*t_o*sqrt(c),
                   sd = sqrt(s*(c + d*c) + 1 + d*c),
                   lower.tail = ifelse(direction == 1, FALSE, TRUE))
    return(power)
  }

  ## Find required relative sample size (upper limit c = 100)
  c <- try(uniroot(f = function(c) powerFun(c) - power,
                   lower = 0, upper = 100)$root)
  if (class(c) == "try-error") return(NA)
  return(c)
}

# ============================================================================
# Parameters of gaussian predictive distributions
```

```r
# ============================================================================
# Default value for tau = 0.08, corresponding to difference of 0.975 and
# 0.025 quantiles of effects on correlation scale with zero mean being the same
# as a medium effect size (r = 0.3), using the classification from Cohen (1992)
tau_default <- 0.08

predictive_flatprior_params <- function(theta_o, sigma_o, sigma_r,
                                        tau = tau_default) {
  mu <- theta_o
  sigma <- sqrt(sigma_o^2 + sigma_r^2 + 2*tau^2)
  return(data.frame("mu" = mu, "sigma" = sigma))
}

predictive_shrinkageprior_params <- function(theta_o, sigma_o, sigma_r,
                                             tau = tau_default) {
  z <- theta_o/sigma_o
  s <- pmax(1 - 1/z^2 - tau^2/theta_o^2, 0)
  mu <- s*theta_o
  sigma <- sqrt(s*(sigma_o^2 + tau^2) + sigma_r^2 + tau^2)
  return(data.frame("mu" = mu, "sigma" = sigma))
}

params_methods_list <- list(
  "Predictive" = function(theta_o, sigma_o, sigma_r, tau) {
  predictive_flatprior_params(theta_o, sigma_o, sigma_r, tau = 0)
  },
  "PredictiveHeterogeneity" = predictive_flatprior_params,
  "Shrinkage" = function(theta_o, sigma_o, sigma_r, tau) {
    predictive_shrinkageprior_params(theta_o, sigma_o, sigma_r, tau = 0)
  },
  "ShrinkageHeterogeneity" = predictive_shrinkageprior_params
  )

params_predictive_distribution <- function(theta_o, sigma_o, sigma_r,
                                           tau = tau_default,
                                           methods = params_methods_list) {
  params <- lapply(methods, function(f) f(theta_o = theta_o, sigma_o = sigma_o,
                                          sigma_r = sigma_r, tau = tau))
  return(params)
}

# ============================================================================
# Binary prediction based on parameters of predictive gaussian distribution
# ============================================================================
significance_gaussian_params <- function(mu, sigma, sigma_r, alpha = 0.05) {
  direction <- ifelse(mu >= 0, 1, -1)
  p_significant <- pnorm(direction*qnorm(1 - alpha/2), mean = mu/sigma_r,
                         sd = sigma/sigma_r,
                         lower.tail = ifelse(direction == 1, FALSE, TRUE))
  return(p_significant)
}
```

```r
predict_significance_params <- function(theta_o, sigma_o, sigma_r,
                                        tau = tau_default, alpha = 0.05) {
  params <- params_predictive_distribution(theta_o = theta_o, sigma_o = sigma_o,
                                           sigma_r = sigma_r, tau = tau)
  p_significant <- lapply(params, function(method) {
    significance_gaussian_params(mu = method$mu, sigma = method$sigma,
                                 sigma_r = sigma_r, alpha = alpha)
    })
  return(p_significant)
}

# ============================================================================
# PIT values of replication studies outcomes
# ============================================================================
pit <- function(theta_o, theta_r, sigma_o, sigma_r, tau = tau_default) {
  params <- params_predictive_distribution(theta_o = theta_o, sigma_o = sigma_o,
                                           sigma_r = sigma_r, tau = tau)
  pit_values <- lapply(params, function(method) {
    pnorm(q = theta_r, mean = method$mu, sd = method$sigma)
    })
  return(pit_values)
}

# ============================================================================
# Proper scoring rules for gaussian predictive distribution
# (Gneiting & Katzfuss (2014))
# ============================================================================
quadr_score_gaussian <- function(mu, sigma, y) {
  score <- -2/sigma*dnorm(x = (y - mu)/sigma) + 1/(2*sqrt(pi)*sigma)
  return(score)
}

log_score_gaussian <- function(mu, sigma, y) {
  score <- (y - mu)^2/(2*sigma^2) + log(sigma) + 0.5*log(2*pi)
  return(score)
}

crp_score_gaussian <- function(mu, sigma, y) {
  score <- sigma*((y - mu)/sigma*(2*pnorm(q = (y - mu)/sigma) - 1) +
                  2*dnorm(x = (y - mu)/sigma) - 1/sqrt(pi))
  return(score)
}

ds_score_gaussian <- function(mu, sigma, y) {
  score <- (y - mu)^2/sigma^2 + 2*log(sigma)
  return(score)
}

gaussian_scoring_rules <- list("QS" = quadr_score_gaussian,
                               "LS" = log_score_gaussian,
```

```
                                "CRPS" = crp_score_gaussian,
                                "DSS" = ds_score_gaussian)

gaussian_scores <- function(theta_o, theta_r, sigma_o, sigma_r,
                            tau = tau_default,
                            scoring_rules = gaussian_scoring_rules) {
  params <- params_predictive_distribution(theta_o = theta_o, sigma_o = sigma_o,
                                           sigma_r = sigma_r, tau = tau)
  results_list <- lapply(seq(scoring_rules), function(i) {
    scores <- lapply(params, function(method) {
      scoring_rules[[i]](mu = method$mu, sigma = method$sigma, y = theta_r)
    })
    data.frame(scores, Type = names(scoring_rules)[i])
  })
  results_df <- do.call(rbind, results_list)
  return(results_df)
}

# ============================================================================
# Proper scoring rules for binary prediction (Held (2014))
# ============================================================================
brier_score <- function(y, p) {
  score <- (y - p)^2
  return(score)
}

binary_scoring_rules <- list("BS" = brier_score)

binary_scores_params <- function(theta_o, sigma_o, sigma_r, tau = tau_default,
                                 pval_r, alpha = 0.05,
                                 scoring_rules = binary_scoring_rules) {
  y <- as.integer(pval_r < alpha)
  p <- predict_significance_params(theta_o = theta_o, sigma_o = sigma_o,
                                   sigma_r = sigma_r, tau = tau)
  p_df <- data.frame(p)
  results_list <- lapply(seq(scoring_rules), function(i) {
    scores <- scoring_rules[[i]](y = y, p = p_df)
    data.frame(scores, "Type" = names(scoring_rules)[i])
  })
  results_df <- do.call(rbind, results_list)
  return(results_df)
}

# ============================================================================
# (Mis-) Calibration tests (Held, Rufibach, Balabdaoui (2010))
# ============================================================================
# Note: LS an DSS Scoring rules without constant terms were used in paper,
# whereas the definitions in Gneiting & Katzfuss involve also constants

# Unconditional miscalibration tests
score_calib_test <- function(theta_o, theta_r, sigma_o, sigma_r,
```

```
                              tau = tau_default) {
  # scoring rules according to definitions in paper
  scoring_rules_test <- list("LS" = function(mu, sigma, y) {
    0.5*(log(sigma^2) + ((y - mu)/sigma)^2)
  },
                             "CRPS" = crp_score_gaussian)
  scores <- gaussian_scores(theta_o = theta_o, theta_r = theta_r,
                            sigma_o = sigma_o, sigma_r = sigma_r,
                            tau = tau, scoring_rules = scoring_rules_test)
  params <- params_predictive_distribution(theta_o = theta_o, sigma_o = sigma_o,
                                           sigma_r = sigma_r, tau = tau)
  n <- length(theta_o)

  test_log <- lapply(seq(ncol(scores) - 1), function(i) {
    mean_LS <- mean(scores[scores$Type == "LS",i])
    expectation <- 0.5 + mean(log(params[[i]]$sigma))
    variance <- 1/(2*n)
    test_statistic <- (mean_LS - expectation)/sqrt(variance)
    test_pvalue <- 2*pnorm(q = abs(test_statistic), lower.tail = FALSE)
    data.frame("t" = test_statistic, "pvalue" = test_pvalue, "Test" = "LS",
               "Method" = colnames(scores)[i])
  })
  test_crps <- lapply(seq(ncol(scores) - 1), function(i) {
    mean_CRPS <- mean(scores[scores$Type == "CRPS",i])
    expectation <- 1/sqrt(pi)*mean(params[[i]]$sigma)
    variance <- 0.1627516/n^2 * sum(params[[i]]$sigma^2)
    test_statistic <- (mean_CRPS - expectation)/sqrt(variance)
    test_pvalue <- 2*pnorm(q = abs(test_statistic), lower.tail = FALSE)
    data.frame("t" = test_statistic, "pvalue" = test_pvalue, "Test" = "CRPS",
               "Method" = colnames(scores)[i])
  })
  tests_df <- do.call(rbind, c(test_log, test_crps))
  return(tests_df)
}

# Score regression calibration tests
# 1) DSS_i = a + b*log(sigma_i) + e_i
# ===> e_i homoscedastic, H0: a = 0.5, b = 1
# 2) CRPS_i = c + d*sigma_i + e_i
# ===> e_i heteroscedastic (w_i = 1/sigma^2_i), H0: c = 0, d = 1/sqrt(pi)
score_calib_regr_test <- function(theta_o, theta_r, sigma_o, sigma_r,
                                  tau = tau_default) {
  # scoring rules according to definitions in paper
  scoring_rules_test <- list("DSS" = function(mu, sigma, y) {
    0.5*(log(sigma^2) + ((y - mu)/sigma)^2)
  },
                             "CRPS" = crp_score_gaussian)
  scores <- gaussian_scores(theta_o = theta_o, theta_r = theta_r,
                            sigma_o = sigma_o, sigma_r = sigma_r,
                            tau = tau, scoring_rules = scoring_rules_test)
  params <- params_predictive_distribution(theta_o = theta_o, sigma_o = sigma_o,
```

```
                                          sigma_r = sigma_r, tau = tau)

  test_dss <- lapply(seq(ncol(scores) - 1), function(i) {
    dss_i <- scores[scores$Type == "DSS",i]
    sigma_i <- params[[i]]$sigma
    fit_dss_i <- lm(dss_i ~ 1 + log(sigma_i))
    ab_diff_i <- matrix(coef(fit_dss_i) - c(0.5, 1))
    statistic <- t(ab_diff_i) %*% solve(vcov(fit_dss_i)) %*% ab_diff_i
    test_pvalue <- pchisq(q = statistic, 2, lower.tail = FALSE)
    data.frame("t" = statistic, "pvalue" = test_pvalue,
               "Test" = "DSS-Regression", "Method" = colnames(scores)[i])
  })
  test_crps <- lapply(seq(ncol(scores) - 1), function(i) {
    crps_i <- scores[scores$Type == "CRPS",i]
    sigma_i <- params[[i]]$sigma
    fit_crps_i <- lm(crps_i ~ 1 + sigma_i, weights = 1/sigma_i^2)
    cd_diff_i <- matrix(coef(fit_crps_i) - c(0, 1/sqrt(pi)))
    statistic <- t(cd_diff_i) %*% solve(vcov(fit_crps_i)) %*% cd_diff_i
    test_pvalue <- pchisq(q = statistic, 2, lower.tail = FALSE)
    data.frame("t" = statistic, "pvalue" = test_pvalue,
               "Test" = "CRPS-Regression", "Method" = colnames(scores)[i])
  })
  tests_df <- do.call(rbind, c(test_dss, test_crps))
  return(tests_df)
}

# Spiegelhalter's z-statistic (Spiegelhalter, 1986)
brier_z_test_params <- function(theta_o, sigma_o, sigma_r, tau = tau_default,
                                pval_r, alpha = 0.05) {
  p <- predict_significance_params(theta_o = theta_o, sigma_o = sigma_o,
                                   sigma_r = sigma_r, tau = tau, alpha = alpha)
  y <- as.integer(pval_r < alpha)
  tests <- lapply(seq(p), function(i) {
    z <- sum((y - p[[i]])*(1 - 2*p[[i]]))/sqrt(sum((1 - 2*p[[i]])^2*
                                                     p[[i]]*(1 - p[[i]])))
    pval <- 2*pnorm(q = abs(z), lower.tail = FALSE)
    data.frame("z" = z, "pvalue" = pval, "Method" = names(p)[i])
  })
  tests_df <- do.call(rbind, tests)
  return(tests_df)
}

# ============================================================================
# Area und the curve (AUC)
# ============================================================================
auc_analysis_params <-  function(theta_o, sigma_o, sigma_r, tau = tau_default,
                                 pval_r, alpha = 0.05) {
  p <- predict_significance_params(theta_o = theta_o, sigma_o = sigma_o,
                                   sigma_r = sigma_r, tau = tau, alpha = alpha)
  y <- as.integer(pval_r < alpha)
  ind_cases <- y == 1
```

```
  result <- lapply(names(p), function(method) {
    auc <- biostatUZH::confIntAUC(cases = p[[method]][ind_cases],
                                  controls = p[[method]][!ind_cases],
                                  conf.level = 1 - alpha)
    data.frame("CI lower" = auc$lower[2],
               "AUC" = auc$AUC[2],
               "CI upper" = auc$upper[2],
               "Method" = method)
  })
  result_df <- do.call(rbind, result)
  return(result_df)
}

# ============================================================================
# Calibration slope
# ============================================================================
calib_slope_continuous <- function(theta_o, theta_r, sigma_o, sigma_r,
                                   tau = tau_default) {
  params <- params_predictive_distribution(theta_o = theta_o, sigma_o = sigma_o,
                                           sigma_r = sigma_r, tau = tau)
  result <- lapply(names(params), function(method) {
    tmp_data <- data.frame(y = theta_r,
                           yhat = params[[method]]$mu,
                           v = sigma_r^2 + sigma_o^2)
    fit <- lm(y ~ yhat, data = tmp_data)
    slope_ci <- confint(fit)
    data.frame("CI lower" = slope_ci[2,1], "Slope" = unname(coef(fit)[2]),
               "CI upper" = slope_ci[2,2], "Method" = method)
  })
  result_df <- do.call(rbind, result)
  return(result_df)
}

calib_slope_binary_params <- function(theta_o, sigma_o, sigma_r,
                                      tau = tau_default, pval_r,
                                      alpha = 0.05) {
  p <- predict_significance_params(theta_o = theta_o, sigma_o = sigma_o,
                                   sigma_r = sigma_r, tau = tau, alpha = alpha)
  y <- as.integer(pval_r < alpha)

  result <- lapply(names(p), function(method) {
    tmp_data <- data.frame(p = p[[method]],
                           logit_p = qlogis(p = p[[method]]),
                           y = y)
    logist_fit <- try(glm(y ~ logit_p, family = "binomial", data = tmp_data))
    if(inherits(logist_fit, "try-error")) {
      NA_df <- data.frame("CI lower" = NA, "Slope" = NA,
                          "CI upper" = NA, "Method" = method)
      return(NA_df)
    } else {
```

```
      slope_ci <- confint.default(logist_fit)
      data.frame("CI lower" = slope_ci[2,1],
                 "Slope" = unname(coef(logist_fit)[2]),
                 "CI upper" = slope_ci[2,2], "Method" = method)
    }
  })
  result_df <- do.call(rbind, result)
  return(result_df)
}

# ============================================================================
# Expected vs. observed number of statistically significant replications
# ============================================================================
expected_significant_params <- function(theta_o, sigma_o, sigma_r,
                                        tau = tau_default,
                                        pval_r, alpha = 0.05) {
  n <- length(sigma_o)
  observed <- sum(as.integer(pval_r < alpha))
  p <- predict_significance_params(theta_o = theta_o, sigma_o = sigma_o,
                                   sigma_r = sigma_r, tau = tau, alpha = alpha)
  expected <- sapply(p, sum)
  result <- lapply(seq(length(p)), function(i) {
    X <- (observed - expected[i])^2/expected[i] +
      ((n - observed) - (n - expected[i]))^2/(n - expected[i])
    pval <- pchisq(q = X, df = 1, lower.tail = FALSE)
    data.frame("N" = n, "Observed" = observed, "Expected" = expected[i],
               "pvalue" = pval, "Method" = names(p)[i])
  })
  result_df <- do.call(rbind, result)
  return(result_df)
}

# ============================================================================
# Prediction intervals
# ============================================================================
prediction_intervals <- function(theta_o, sigma_o, sigma_r, tau = tau_default,
                                 gamma = 0.95) {
  params <- params_predictive_distribution(theta_o = theta_o, sigma_o = sigma_o,
                                           sigma_r = sigma_r, tau = tau)
  result <- lapply(seq(params), function(i) {
    data.frame("PI lower" = qnorm(p = (1 - gamma)/2, mean = params[[i]]$mu,
                                  sd = params[[i]]$sigma),
               "y hat" = params[[i]]$mu,
               "PI upper" = qnorm(p = (1 + gamma)/2, mean = params[[i]]$mu,
                                  sd = params[[i]]$sigma),
               "Method" = names(params)[i])
  })
  result_df <- do.call(rbind, result)
 return(result_df)
}
```

## A.2 Data preprocessing

```
# ============================================================================
# Reproducibility project Psychology (rpp)
# ============================================================================
library(tidyverse)
url_master <- "https://github.com/CenterForOpenScience/rpp/archive/master.zip"
download.file(url = url_master, destfile = "rpp_git_repo.zip")
download.file(url = "https://osf.io/kn7f4/download",
              destfile = "rpp_prediction_markets.csv")
unzip("rpp_git_repo.zip" )
rpp_prediction_market <- read.csv("rpp_prediction_markets.csv")

# ATTENTION: if running on linux: comment out the windows command
# "choose.dir" at beginning of masterscript.R !
setwd("rpp-master/")
source("masterscript.R")
setwd("../")

MASTER_cleaned <- MASTER %>%
  mutate(FZ_OS = atanh(T_r..O.),
         FZ_RS = atanh(T_r..R.),
         Actual.Power..O. = as.double(as.character(Actual.Power..O.)),
         Power..R. = as.double(Power..R.)) %>%
  select(ID, Authors..O., Journal..O.,  Discipline..O.,
         T_Test.Statistic..O., T_df1..O., T_df2..O., T_Test.value..O.,
         Type.of.analysis..O., Effect.size..O., Actual.Power..O., T_N..O.,
         T_r..O., T_pval_USE..O., FZ_OS, T_Test.Statistic..R., T_df1..R.,
         T_df2..R., T_Test.value..R., Type.of.analysis..R., Power..R.,
         Effect.Size..R., T_N..R., T_r..R., T_pval_USE..R., FZ_RS,
         Meta.analytic.estimate..Fz.) %>%
  rename(Study_ID = ID, Authors_OS = Authors..O., Journal_OS = Journal..O.,
         Discipline = Discipline..O.,
         Type_Test_Statistic_OS = T_Test.Statistic..O., DF1_OS = T_df1..O.,
         DF2_OS = T_df2..O., Test_Statistic_OS = T_Test.value..O.,
         N_OS = T_N..O., r_OS = T_r..O., pval_OS = T_pval_USE..O.,
         Analysis_Type_OS = Type.of.analysis..O.,
         Effect_Size_OS = Effect.size..O., Power_OS = Actual.Power..O.,
         Type_Test_Statistic_RS = T_Test.Statistic..R., DF1_RS = T_df1..R.,
         DF2_RS = T_df2..R., Test_Statistic_RS = T_Test.value..R.,
         N_RS = T_N..R., r_RS = T_r..R., pval_RS = T_pval_USE..R.,
         Analysis_Type_RS = Type.of.analysis..R.,
         Effect_Size_RS = Effect.Size..R., Power_RS = Power..R.,
         FZ_meta = Meta.analytic.estimate..Fz.)

MASTER_cleaned$FZ_se_OS[!is.na(MASTER_cleaned$FZ_meta)] <- final$sei.o
MASTER_cleaned$FZ_se_RS[!is.na(MASTER_cleaned$FZ_meta)] <- final$sei.r
MASTER_cleaned %>%
  mutate(pval_OS = ifelse(Study_ID %in% c(7, 15, 47, 94, 120, 140),
                          pval_OS*2, pval_OS), # these were one-sided p-values
         pval_RS = ifelse(Study_ID %in% c(7, 15, 47, 94, 120, 140),
```

```
                            pval_RS*2, pval_RS), # according supplementary
         pvalFZ_OS = 2*pnorm(FZ_OS/FZ_se_OS, lower.tail = FALSE),
         pvalFZ_RS = 2*pnorm(FZ_RS/FZ_se_RS, lower.tail = FALSE))
write_csv(MASTER_cleaned, path = "RPP.csv")

# ============================================================================
# Experimental economics replication project (eerp)
# ============================================================================
# - All files downloaded from https://osf.io/pnwuz/
# - Unfortunately this data had to be manually recorded from the file
#   "create_studydetails.do" since I do not own the commercial software stata
#   which is required to run the .do file and generate .dat file (economists ..)
# - Similarly, effective sample size taken from  "effectstandardization.py" file
# - Prediction market infos taken from table S3 in Supplementary of Article
#   http://science.sciencemag.org/content/sci/suppl/2016/03/02/
#   science.aaf0918.DC1/aaf0918-Camerer-SM.pdf

library(tidyverse)
Study <- c("Abeler et al. (AER 2011)",
           "Ambrus and Greiner (AER 2012)",
           "Bartling et al. (AER 2012)",
           "Charness and Dufwenberg (AER 2011)",
           "Chen and Chen (AER 2011)",
           "de Clippel et al. (AER 2014)",
           "Duffy and Puzzello (AER 2014)",
           "Dulleck et al. (AER 2011)",
           "Fehr et al. (AER 2013)",
           "Friedman and Oprea (AER 2012)",
           "Fudenberg et al. (AER 2012)",
           "Huck et al. (AER 2011)",
           "Ifcher and Zarghamee (AER 2011)",
           "Kessler and Roth (AER 2012)",
           "Kirchler et al (AER 2012)",
           "Kogan et al. (AER 2011)",
           "Kuziemko et al. (QJE 2014)",
           "Ericson and Fuster (QJE 2011)")
Market_Belief <- c(0.696, 0.692, 0.805, 0.695, 0.778, 0.759, 0.806, 0.738, 0.629,
                   0.833, 0.933, 0.920, 0.588, 0.937, 0.712, 0.802, 0.632, 0.622)
Survey_Belief_premarket <- c(0.696, 0.542, 0.807, 0.715, 0.682, 0.730, 0.685,
                             0.807, 0.674, 0.863, 0.790, 0.749, 0.542, 0.837,
                             0.704, 0.748, 0.568, 0.658)
Survey_Belief_postmarket <- c(0.697, 0.620, 0.733, 0.708, 0.692, 0.716, 0.694,
                              0.744, 0.666, 0.817, 0.770, 0.730, 0.566, 0.825,
                              0.728, 0.752, 0.582, 0.650)
pval_OS <- c(0.046, 0.057, 0.007, 0.01, 0.033, 0.001, 0.01, 0.0001, 0.011,
             4*10^(-11), 0.001, 0.0039, 0.031, 1.631*10^(-18), 0.0163, 0.000026,
             0.07, 0.03)
# session numbers, not effective sample size!
# N_OS <- c(120, 117, 216, 162, 72, 158, 54, 168, 60, 78, 124,120, 58, 288, 120,
#           126, 42, 112)
# these are effective sample size:
```

```
N_OS <- c(120, 39, 12, 43, 6, 790, 9, 21, 30, 78, 124, 12, 58, 288, 12, 160,
          42, 104)
r_OS <- c(0.182821975588, 0.310518647505, 0.719849875686, 0.383943377571,
          0.842508557739, 0.117768981826, 0.761510174904, 0.722509234548,
          0.453281944406, 0.642590125822, 0.303741608956, 0.832065702534,
          0.282103220856, 0.486223364603, 0.664409308738, 0.323896842962,
          0.282261740933, 0.212921823047)
pval_RS <- c(0.16, 0.012, 0.001, 0.003, 0.571, 4*10^(-10), 0.674, 0.0008,
             0.026, 0.004276, 0.0001506473, 0.1415, 0.933, 0.016, 0.0095,
             0.001, 0.154, 0.0546)
# session numbers, not effective sample size!
# N_RS <- c(318, 357, 360, 264, 168, 156, 96, 128, 102, 40, 128, 160, 131, 48,
#           220, 90, 144, 262)
# these are effective sample size:
N_RS <- c(318, 119, 20, 65, 14, 780, 16, 16, 51, 40, 128, 16, 131, 48, 22,
          112, 144, 248)
r_RS <- c(0.0790703532018, 0.229536356959, 0.657411288278, 0.363002684809,
          0.170189166838, 0.266538269195, -0.11596300548, 0.731605727911,
          0.311199311719, 0.437953607707, 0.326573539422, 0.367593525514,
          -0.00701629144787, 0.34463841128, 0.533556821497, 0.304223231247,
          -0.119848901099, 0.122871403263)
Power_RS <- c(0.9, 0.91, 0.94, 0.9, 0.9, 0.9, 0.93, 0.92, 0.91, 0.99, 0.92,
              0.91, 0.9, 0.95, 0.9, 0.94, 0.92, 0.91)

tibble(Study, r_OS, N_OS, pval_OS, r_RS, N_RS, pval_RS, Power_RS,
       Market_Belief, Survey_Belief_premarket, Survey_Belief_postmarket) %>%
  mutate(FZ_OS = atanh(r_OS),
         FZ_se_OS = 1/sqrt(N_OS - 3),
         pvalFZ_OS = 2*pnorm(FZ_OS/FZ_se_OS, lower.tail = FALSE),
         pvalFZ_OS = ifelse(pvalFZ_OS > 1, 1, pvalFZ_OS),
         FZ_RS = atanh(r_RS),
         FZ_se_RS = 1/sqrt(N_RS - 3),
         pvalFZ_RS = 2*pnorm(FZ_RS/FZ_se_RS, lower.tail = FALSE),
         pvalFZ_OS = ifelse(pvalFZ_RS > 1, 1, pvalFZ_RS)) %>%
  write_csv(path = "EERP.csv")

# ============================================================================
# Social sciences replication project (ssrp)
# ============================================================================
library(tidyverse)
names <- c("Ackerman et al. (2010), Science",
           "Aviezer et al. (2012), Science",
           "Balafoutas and Sutter (2012), Science",
           "Derex et al. (2013), Nature",
           "Duncan et al. (2012), Science",
           "Gervais and Norenzayan (2012), Science",
           "Gneezy et al. (2014), Science",
           "Hauser et al. (2014), Nature",
           "Janssen et al. (2010), Science",
           "Karpicke and Blunt (2011), Science",
           "Kidd and Castano (2013), Science",
```

```
            "Kovacs et al. (2010), Science",
            "Lee and Schwarz (2010), Science",
            "Morewedge et al. (2010), Science",
            "Nishi et al. (2015), Nature",
            "Pyc and Rawson (2010), Science",
            "Ramirez and Beilock (2011), Science",
            "Rand et al. (2012), Nature",
            "Shah et al. (2012), Science",
            "Sparrow et al. (2011), Science",
            "Wilson et al. (2014), Science")
download.file(url = "https://osf.io/abu7k/download",
              destfile = "SSRP_Data_Processed.csv")
download.file(url = "https://osf.io/vr6p8/download",
              destfile = "SSRP_Data_Peer_Beliefs_Processed.csv")
ssrp <- read_csv("SSRP_Data_Processed.csv")
ssrp_pmarket <- read_csv("SSRP_Data_Peer_Beliefs_Processed.csv")
prediction_markets <- ssrp_pmarket %>%
  select(m3_p, m3_b) %>%
  rename(Market_Belief = m3_p,
         Survey_Belief = m3_b) %>%
  filter(!is.na(Market_Belief))

ssrp %>%
  mutate(Name_OS = names,
         FZ_OS = atanh(r_os),
         FZ_se_OS = 1/sqrt(n_os - 3),
         FZ_RS1 = atanh(r_rs1),
         FZ_se_RS1 = 1/sqrt(n_rs1 - 3),
         FZ_RS2 = atanh(r_rs2),
         FZ_se_RS2 = 1/sqrt(n_rs2 - 3)) %>%
  select(study, Name_OS, sref, type_os, stat_os, n_os, in_os, r_os, r95l_os,
         r95u_os, p_os, FZ_OS, FZ_se_OS, type_rs1, stat_rs1, n_rs1, in_rs1,
         r_rs1, r95l_rs1, r95u_rs1, p_rs1, pow_rs1, FZ_RS1, FZ_se_RS1,
         type_rs2, stat_rs2, n_rs2, in_rs2, r_rs2, r95l_rs2,
         r95u_rs2, p_rs2, pow_rs2, FZ_RS2, FZ_se_RS2) %>%
  rename(Study = study, Sref = sref, Type_OS = type_os, Stat_OS = type_os,
         N_OS = n_os, In_OS = in_os, r_OS = r_os, r95l_OS = r95l_os,
         r95u_OS = r95u_os, pval_OS = p_os, Type_RS1 = type_rs1,
         Stat_RS1 = stat_rs1, N_RS1 = n_rs1, In_RS1 = in_rs1, r_RS1 = r_rs1,
         r95l_RS1 = r95l_rs1, r95u_RS1 = r95u_rs1, pval_RS1 = p_rs1,
         Power_RS1 = pow_rs1, Type_RS2 = type_rs2, Stat_RS2 = stat_rs2,
         N_RS2 = n_rs2, In_RS2 = in_rs2, r_RS2 = r_rs2, r95l_RS2 = r95l_rs2,
         r95u_RS2 = r95u_rs2, pval_RS2 = p_rs2, Power_RS2 = pow_rs2) %>%
  mutate(r_RS = ifelse(!is.na(r_RS2), r_RS2, r_RS1),
         FZ_RS = ifelse(!is.na(FZ_RS2), FZ_RS2, FZ_RS1),
         FZ_se_RS = ifelse(!is.na(FZ_se_RS2), FZ_se_RS2, FZ_se_RS1),
         pval_RS = ifelse(!is.na(FZ_se_RS2), pval_RS2, pval_RS1),
         N_RS = ifelse(!is.na(FZ_se_RS2), N_RS2, N_RS1)) %>%
  bind_cols(., prediction_markets) %>%
  write_csv(path = "SSRP.csv")
```

```r
# ============================================================================
# Experimental philosophy replicability project (rpphi)
# ============================================================================
library(tidyverse)
download.file(url = "https://osf.io/4ewkh/download",
              destfile = "XPhiReplicability_CompleteData.csv")
rpphi <- read.csv("XPhiReplicability_CompleteData.csv", stringsAsFactors = FALSE)
rpphi %>%
  mutate(FZ_OS = atanh(OriginalRES),
         FZ_RS = atanh(ReplicationRES),
         FZ_se_OS = 1/sqrt(OriginalN_Effect - 3),
         FZ_se_RS = 1/sqrt(ReplicationN_Effect - 3),
         pval_RS = 2*pnorm(abs(FZ_RS/FZ_se_RS), lower.tail = FALSE),
         pval_RS = ifelse(pval_RS > 1, 1, pval_RS),
         pval_OS = 2*pnorm(abs(FZ_OS/FZ_se_OS), lower.tail = FALSE),
         pval_OS = ifelse(pval_OS > 1, 1, pval_OS)) %>%
  select(PAPER_ID, OriginalN_Effect, OriginalTEST, OriginalEFFECTSIZE,
         OriginalANALYSIS, OriginalRES, OriginalPOWER, OriginalR95CI,
         FZ_OS, FZ_se_OS, pval_OS, ReplicationN_Effect, ReplicationTEST,
         ReplicationANALYSIS, ReplicationEFFECTSIZE, ReplicationRES,
         ReplicationR95CI, FZ_RS, FZ_se_RS, pval_RS, ReplicationSUCCESS, OSF) %>%
  rename(Study = PAPER_ID, Type_Test_OS = OriginalTEST,
         Test_Statistic_OS = OriginalANALYSIS, N_OS = OriginalN_Effect,
         r_OS = OriginalRES, r_CI_OS = OriginalR95CI,
         Effect_Size_OS = OriginalEFFECTSIZE, Power_OS = OriginalPOWER,
         Type_Test_RS = ReplicationTEST, Test_Statistic_RS = ReplicationANALYSIS,
         N_RS = ReplicationN_Effect, r_RS = ReplicationRES,
         r_CI_RS = ReplicationR95CI, Effect_Size_RS = ReplicationEFFECTSIZE,
         Replication_Success = ReplicationSUCCESS) %>%
 write_csv(path = "RPPHI.csv")

# ============================================================================
# Combining all data sets
# ============================================================================
library(tidyverse)
rpp <- read_csv("RPP/RPP.csv")
rpphi <- read_csv("RPPHI/RPPHI.csv")
ssrp <- read_csv("SSRP/SSRP.csv")
eerp <- read_csv("EERP/EERP.csv")

# Subsets of data where effect sizes transformed to correlations available
rpp_correlations_subset <- rpp %>%
  filter(!is.na(r_OS) & !is.na(r_RS)) %>%
  mutate(Project = "Psychology",
         Study = Authors_OS,
         Survey_Belief_Premarket = NA,
         Survey_Belief_Postmarket = NA,
         Market_Belief = NA) %>%
  select(Study, r_OS, r_RS, FZ_OS, FZ_RS, FZ_se_OS, FZ_se_RS, pval_RS,
         N_OS, N_RS, pval_OS, Project, Survey_Belief_Premarket,
         Survey_Belief_Postmarket, Market_Belief)
```

```
eerp_correlations_subset <- eerp %>%
  mutate(Project = "Experimental Economics",
         Stuy = Study,
         Survey_Belief_Premarket = eerp$Survey_Belief_premarket,
         Survey_Belief_Postmarket = eerp$Survey_Belief_postmarket,
         Market_Belief = eerp$Market_Belief) %>%
  select(Study, r_OS, r_RS, FZ_OS, FZ_RS, FZ_se_OS, FZ_se_RS, pval_RS,
         N_OS, N_RS, pval_OS, Project, Survey_Belief_Premarket,
         Survey_Belief_Postmarket, Market_Belief)

ssrp_correlations_subset <- ssrp %>%
  mutate(Project = "Social Sciences",
         Study = Name_OS,
         Survey_Belief_Premarket = ssrp$Survey_Belief,
         Survey_Belief_Postmarket = NA,
         Market_Belief = ssrp$Market_Belief) %>%
  select(Study, r_OS, r_RS, FZ_OS, FZ_RS, FZ_se_OS, FZ_se_RS, pval_RS,
         N_OS, N_RS, pval_OS, Project, Survey_Belief_Premarket,
         Survey_Belief_Postmarket, Market_Belief)

rpphi_correlations_subset <- rpphi %>%
  filter(!is.na(r_OS) & !is.na(r_RS) & !is.na(pval_RS)) %>%
  mutate(Project = "Experimental Philosophy",
         Study = Study,
         Survey_Belief_Premarket = NA,
         Survey_Belief_Postmarket = NA,
         Market_Belief = NA) %>%
  select(Study, r_OS, r_RS, FZ_OS, FZ_RS, FZ_se_OS, FZ_se_RS, pval_RS,
         N_OS, N_RS, pval_OS, Project, Survey_Belief_Premarket,
         Survey_Belief_Postmarket, Market_Belief)

data_correlations_subset <- rbind(rpp_correlations_subset,
                                  eerp_correlations_subset,
                                  ssrp_correlations_subset,
                                  rpphi_correlations_subset) %>%
  mutate(pval_RS_significant = factor(pval_RS < 0.05,
                                      labels = c("Not Significant",
                                                 "Significant")))

write_csv(data_correlations_subset,
          path = "Data_Final/replications_correlation_subset.csv")

# Subset of data where standard error of Fisher z-transformed
# correlations available (Meta analytic subset)
data_ma_subset <- data_correlations_subset %>%
  filter(!is.na(FZ_se_OS) & !is.na(FZ_se_RS))

write_csv(data_ma_subset, path = "Data_Final/replications_ma_subset.csv")
```

```
sessionInfo()

## R version 3.6.1 (2019-07-05)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.3 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
##
## 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
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] tables_0.8.7       Hmisc_4.2-0        Formula_1.2-3
##  [4] survival_2.43-3    lattice_0.20-38    forcats_0.4.0
##  [7] stringr_1.4.0      dplyr_0.8.0.1      purrr_0.3.2
## [10] readr_1.3.1        tidyr_0.8.3        tibble_2.1.1
## [13] ggplot2_3.1.1.9000 tidyverse_1.2.1    knitr_1.22
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.1           lubridate_1.7.4     assertthat_0.2.1
##  [4] digest_0.6.18        plyr_1.8.4          R6_2.4.0
##  [7] cellranger_1.1.0     backports_1.1.4     acepack_1.4.1
## [10] evaluate_0.13        httr_1.4.0          highr_0.8
## [13] pillar_1.3.1         rlang_0.3.4         lazyeval_0.2.2
## [16] readxl_1.3.1         rstudioapi_0.10     data.table_1.12.2
## [19] rpart_4.1-15         Matrix_1.2-17       checkmate_1.9.1
## [22] labeling_0.3         splines_3.6.1       foreign_0.8-70
## [25] htmlwidgets_1.3      biostatUZH_1.8.0    munsell_0.5.0
## [28] broom_0.5.2          compiler_3.6.1      modelr_0.1.4
## [31] xfun_0.6             pkgconfig_2.0.2     base64enc_0.1-3
## [34] htmltools_0.3.6      nnet_7.3-12         tidyselect_0.2.5
## [37] gridExtra_2.3        htmlTable_1.13.1    codetools_0.2-16
## [40] viridisLite_0.3.0    crayon_1.3.4        withr_2.1.2
## [43] grid_3.6.1           nlme_3.1-140        jsonlite_1.6
## [46] gtable_0.3.0         magrittr_1.5        scales_1.0.0
## [49] cli_1.1.0            stringi_1.4.3       reshape2_1.4.3
## [52] latticeExtra_0.6-28 xml2_1.2.0          generics_0.0.2
## [55] boot_1.3-23          RColorBrewer_1.1-2  tools_3.6.1
## [58] glue_1.3.1           hms_0.4.2           colorspace_1.4-1
## [61] cluster_2.1.0        rvest_0.3.3         haven_2.1.0
```

# Bibliography


Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., Boeck, P. D., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Ho, T. H., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Zandt, T. V., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour*, **2**, 6 – 10. 35

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. John Wiley & Sons, Inc. 42

Camerer, C., 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. 2, 3, 6, 21, 32, 44, 46, 47

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. 2, 3, 6, 21, 32, 44, 46, 47

Cohen, J. (1992). A power primer. *Psychological Bulletin*, **112**, 155 – 159. 15, 16, 45

Cooper, H., Hedges, V., and Valentine, J. C. (2009). *The Handbook of Research Synthesis and Meta-Analysis*. Russel Sage Foundation New York. 6

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society*, **45**, 311 – 354. 8

Copas, J. B. (1997). Using regression models for prediction: shrinkage and regression to the mean. *Statistical Methods in Medical Research*, **6**, 167 – 183. 9

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai van Dongen, 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., Liao, S.-y., 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. 2, 3, 6, 22, 32, 44, 46

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, **45**, 562 – 565. 17

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. 21, 47

Dwan, K., Gamble, C., Williamson, P. R., and and, J. J. K. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias — an updated review. *PLoS ONE*, **8**, e66844. 3

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., Davis, W. E., Devos, T., Fletcher, M. M., German, K., Grahe, J. E., Hermann, A. D., Hicks, J. A., Honeycutt, N., Humphrey, B., Janus, M., Johnson, D. J., Joy-Gaba, J. A., Juzeler, H., Keres, A., Kinney, D., Kirshenbaum, J., Klein, R. A., Lucas, R. E., Lustgraaf, C. J., Martin, D., Menon, M., Metzger, M., Moloney, J. M., Morse, P. J., Prislin, R., Razza, T., Re, D. E., Rule, N. O., Sacco, D. F., Sauerberger, K., Shrider, E., Shultz, M., Siemsen, C., Sobocko, K., Sternglanz, R. W., Summerville, A., Tskhay, K. O., van Allen, Z., Vaughn, L. A., Walker, R. J., Weinberg, A., Wilson, J. P., Wirth, J. H., Wortman, J., and Nosek, B. A. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, **67**, 68 – 82. 2, 32, 47

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, **4**, e5738. 3

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, **1**, 3 – 32. 6

Funder, D. C. and Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, **2**, 156–168. 45

Gelman, A. and Loken, E. (2014). The statistical crisis in science. *American Scientist*, **102**, 460 – 465. 2

Gilbert, D. T., King, G., Pettigrew, S., and Wilson, T. D. (2016). Comment on "estimating the reproducibility of psychological science". *Science*, **351**, 1037 – 1040. 3, 10

Gneiting, T., Balabdaoui, F., and Raftery, E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society*, **69**, 243 – 268. 16, 17, 18

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, **1**, 125 – 151. 3, 16, 19, 41, 44

Good, I. J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association*, **53**, 799 – 813. 32

Goodman, S. (1992). A comment on replication, *P*-values and evidence. *Statistics in Medicine*, **11**, 875 – 879. 2, 13

Held, L. (2019a). A new standard for the analysis and design of replication studies. 2, 3, 47

Held, L. (2019b). On the Bayesian interpretation of the harmonic mean $p$-value. *PNAS*, **116**, 5855 – 5856. 32

Held, L., Rufibach, K., and Balabdaoui, F. (2010). A score regression approach to assess calibration of continuous probabilistic predictions. *Biometrics*, **66**, 1295 – 1305. 19, 20

Held, L. and Sabanés Bové, D. (2014). *Applied Statistical Inference*. Springer-Verlag Berlin Heidelberg. 4, 7, 17, 19

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, **2**, e124. 2

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, **23**, 524 – 532. 3

Kicinski, M., Springate, D. A., and Kontopantelis, E. (2015). Publication bias in meta-analyses from the cochrane database of systematic reviews. *Statistics in Medicine*, **34**, 2781 – 2793. 3

Killeen, P. (2006). An alternative to null-hypothesis significance tests. *Psychological Science*, **16**, 345 – 353. 2

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, v., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJzerman, H., John, M.-S., Joy-Gaba, J. A., Barry Kappes, H., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van 't Veer, A. E., Ann Vaughn, L., Vranka, M., Wichman, A. L., Woodzicka, J. A., and Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, **45**, 142 – 152. 2, 32, 47

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, J., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Štěpán Bahník, Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., Rédei, A. C., Cai, H., Cambier, F., Cantarero, K., Carmichael, C. L., Ceric, F., Chandler, J., Chang, J.-H., Chatard, A., Chen, E. E., Cheong, W., Cicero, D. C., Coen, S., Coleman, J. A., Collisson, B., Conway, M. A., Corker, K. S., Curran, P. G., Cushman, F., Dagona, Z. K., Dalgar, I., Rosa, A. D., Davis, W. E., de Bruijn, M., Schutter, L. D., Devos, T., de Vries, M., Doğulu, C., Dozo, N., Dukes, K. N., Dunham, Y., Durrheim, K., Ebersole, C. R., Edlund, J. E., Eller, A., English, A. S., Finck, C., Frankowska, N., Ángel Freyre, M., Friedman, M., Galliani, E. M., Gandi, J. C., Ghoshal, T., Giessner, S. R., Gill, T., Gnambs, T., Ángel Gómez, González, R., Graham, J., Grahe, J. E., Grahek, I., Green, E. G. T., Hai, K., Haigh, M., Haines, E. L., Hall, M. P., Heffernan, M. E., Hicks, J. A., Houdek, P., Huntsinger, J. R., Huynh, H. P., Ijzerman, H., Inbar, Y., Åse H. Innes-Ker, Jiménez-Leal, W., John, M.-S., Joy-Gaba, J. A., Kamiloğlu, R. G., Kappes, H. B., Karabati, S., Karick, H., Keller, V. N., Kende, A., Kervyn, N., Knežević, G., Kovacs, C., Krueger, L. E., Kurapov, G., Kurtz, J., Lakens, D., Lazarević, L. B., Levitan, C. A., Neil A. Lewis, J., Lins, S., Lipsey, N. P., Losee, J. E., Maassen, E., Maitner, A. T., Malingumu, W., Mallett, R. K., Marotta, S. A., Mededović, J., Mena-Pacheco, F., Milfont, T. L., Morris, W. L., Murphy, S. C., Myachykov, A., Neave, N., Neijenhuijs, K., Nelson, A. J., Neto, F., Nichols, A. L., Ocampo, A., O'Donnell, S. L., Oikawa, H., Oikawa, M., Ong, E., Orosz, G., Osowiecka, M., Packard, G., Pérez-Sánchez, R., Petrović, B., Pilati, R., Pinter, B., Podesta, L., Pogge, G., Pollmann, M. M. H., Rutchick, A. M., Saavedra, P., Saeri, A. K., Salomon, E., Schmidt, K., Schönbrodt, F. D., Sekerdej,

M. B., Sirlopú, D., Skorinko, J. L. M., Smith, M. A., Smith-Castro, V., Smolders, K. C. H. J., Sobkow, A., Sowden, W., Spachtholz, P., Srivastava, M., Steiner, T. G., Stouten, J., Street, C. N. H., Sundfelt, O. K., Szeto, S., Szumowska, E., Tang, A. C. W., Tanzer, N., Tear, M. J., Theriault, J., Thomae, M., Torres, D., Traczyk, J., Tybur, J. M., Ujhelyi, A., van Aert, R. C. M., van Assen, M. A. L. M., van der Hulst, M., van Lange, P. A. M., van't Veer, A. E., Vásquez-Echeverría, A., Vaughn, L. A., Vázquez, A., Vega, L. D., Verniers, C., Verschoor, M., Voermans, I. P. J., Vranka, M. A., Welch, C., Wichman, A. L., Williams, L. A., Wood, M., Woodzicka, J. A., Wronska, M. K., Young, L., Zelenski, J. M., Zhijia, Z., and Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, **1**, 443 – 490. 2, 32, 47

Lee, L.-S. (2018). When the alpha is the omega: *P*-values, "substantial evidence," and the 0.05 standard at FDA. *Food and Drug Law Journal*, **72**, 595 – 635. 2

Ly, A., Etz, A., Marsman, M., and Wagenmakers, E.-J. (2018). Replication Bayes factors from evidence updating. 3, 47

McShane, B. B., Tackett, J. L., Böckenholt, U., and Gelman, A. (2019). Large-scale replication projects in contemporary psychological research. *The American Statistician*, **73**, 99 – 105. 3

Murdoch, D. (2018). *tables: Formula-Driven Table Generation*. R package version 0.8.7. 22

Neuenschwander, B., Roychoudhury, S., and Branson, M. (2018). Predictive evidence threshold scaling: Does the evidence meet a confirmatory standard? *Statistics in Biopharmaceutical Research*, **10**, 76 – 84. 15

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, **349**, aac4716. 2, 3, 6, 20, 32, 44

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. 2, 3, 44, 46

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 22

Rufibach, K., Burger, H. U., and Abt, M. (2016). Bayesian predictive power: choice of prior and some recommendations for its use as probability of success in drug development. *Pharmaceutical Statistics*, **15**, 438 – 446. 8

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, **26**, 559 – 569. 2

Spiegelhalter, D. J. (1986). Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, **5**, 421 – 433. 19

Spiegelhalter, D. J., Abrams, R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York: Wiley. 7, 8, 15

Steyerberg, E. (2009). *Clinical Prediction Models*. Springer-Verlag New York. 17

Verhagen, J. and Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, **143**, 1457 – 1475. 3

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1. 22

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with $g$-prior distributions. *Bayesian Inference and Decision techniques*, **6**, 233 – 243. 8