Probabilistic forecasting of replication studies

Samuel Pawel, Leonhard Held
Epidemiology, Biostatistics and Prevention Institute (EBPI)
Center for Reproducible Science (CRS)
University of Zurich, Switzerland
E-mail: samuel.pawel@uzh.ch

December 21, 2022

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 forecasts of the replication outcomes, which we then evaluated regarding discrimination, calibration and sharpness. A novel model, which can take into account both inflation and heterogeneity of effects, was used and predicted 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 was still 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 inflated, possibly caused by publication bias or questionable research practices, and also that some degree of heterogeneity between original and replication effects should be expected. Moreover, the results 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 compatible with significant results from the original study. The developed statistical methods as well as the data sets are available in the `R` package `ReplicationSuccess`.

1 Introduction

Direct replication of past studies is an essential tool in the modern scientific process for assessing the credibility of scientific discoveries. Over the course of the last decade, however, concerns regarding the replicability of scientific discoveries have increased dramatically, leading many to conclude that science is in a crisis (Ioannidis, 2005; Begley and Ioannidis, 2015).

For this reason, researchers in different fields, e. g., psychology or economics, have joined forces to conduct large-scale replication projects. In such a replication project, representative original studies are carefully selected and then direct replication studies of these original studies are carried out. By now, many of the initial projects have been completed and their data made available to the public (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). The low rate of replication success in some of these projects has received enormous attention in the media and scientific 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*).

Making forecasts about an uncertain future is a common human desire and central for decision making in science and society (Gneiting, 2008; Gneiting and Katzfuss, 2014). There have been many attempts to forecast the outcomes of replication studies based on the results from the

original studies (Dreber et al., 2015; Patil et al., 2016; Camerer et al., 2016, 2018; Altmejd et al., 2019; Forsell et al., 2019). This is interesting for various reasons: First, a forecast of how likely a replication will be "successful" according to some criterion (e. g., an effect estimate reaches statistical significance) can help to assess the credibility of the original finding in the first place and inform the decision whether a replication study should be conducted at all. Second, after a replication has been completed, its results can be compared to its forecast in order to assess compatibility between the two findings. Finally, forecasting can also be helpful in designing an informative replication study, for example it can be used for sample size planning.

Although there have been theoretical contributions to the literature long before the replication crisis started (Goodman, 1992; Senn, 2002; Bayarri and Mayoral, 2002), the last years have witnessed new developments regarding forecasting of replication studies. Moreover, due to the increasing popularity of replication studies, forecasts could be evaluated with actual data.

For instance, *prediction markets* have been used in order to estimate the peer belief about whether a replication will result in a statistically significant outcome (Dreber et al., 2015; Camerer et al., 2016, 2018; Forsell et al., 2019). Prediction markets are a tool to aggregate beliefs of market participants regarding the possibility of an investigated outcome and they have been used successfully in numerous domains, e. g., sports and politics. However, despite good predictive performance, taking statistical significance as the target variable of the forecasts requires arbitrary dichotomization of the outcomes, although one would prefer to rather forecast the replication effect estimate itself. Moreover, the evaluation of these forecasts was usually based on ad-hoc measures such as correlation of the estimated probabilities with the outcome. In fields where forecasting is of central importance, e. g., meteorology, climatology, or infectious disease epidemiology, extensive methodology has been developed to specifically assess calibration, discrimination, and sharpness of probabilistic forecasts (Gneiting and Katzfuss, 2014). It is therefore of interest to assess whether more insights about the forecasts can be gained when applying a more state-of-the-art evaluation strategy. Finally, it is also of interest to benchmark the prediction market forecasts with statistical forecasts that do not require recruiting experts and setting up prediction market infrastructure.

A statistical method to obtain probabilistic forecasts of replication estimates was proposed by Patil et al. (2016) and then also used in the analysis of the outcomes of some large-scale replication projects. Specifically, the agreement between the original and replication study was assessed by a prediction interval of the replication effect estimate based on the original effect estimate. This method was illustrated using the data set from the *Reproducibility Project: Psychology* (Open Science Collaboration, 2015), and it 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, the coverage of the 95% prediction intervals was examined to assess predictive performance. Although this evaluation method provides some clue about the calibration of the forecasts, more sophisticated methods exist to assess calibration and sharpness specifically (Gneiting and Katzfuss, 2014). Moreover, the prediction model which was used does not take into account that the original effect estimates may be inflated. In the statistical prediction literature, the phenomenon that future observations of a random quantity tend to be less extreme than the original observation, is commonly known as *regression to the mean* and usually addressed by shrinkage methods (Copas, 1997). This effect might be even more pronounced by the influence of publication bias (Dwan et al., 2013; Kicinski et al., 2015) or questionable research practices (Fanelli, 2009; John et al., 2012). Finally, the model from Patil et al. (2016) also makes the naive assumption that the effect estimates from both studies are realizations of the same underlying effect size, however, there is often between-study heterogeneity (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 the original and replication study.

The objective of this paper is to improve on the previous statistical forecasts and also on their evaluation. In particular, we will develop and evaluate a novel prediction model which can take into account inflation of the original effect estimates as well as between-study heterogeneity of effects. With the available data from large-scale replication projects, we aim to predict the effect estimates of the replication studies based on the estimates from the original studies and knowledge of the sample size in both studies. We will assess the forecasts regarding discrimination, calibration, and sharpness using state-of-the-art evaluation methods from the statistical prediction literature. Finally, we will also benchmark them with the forecasts from prediction markets and the naive model which was used so far.

It is worth pointing out that in our forecasting approach the link between original and replication study is based only on information from the original study and the sample size of the replication study. This is fundamentally different from approaches where the link between original and replication study is estimated from a training sample of past original and replication study pairs, as for example done recently by Altmejd et al. (2019). Since in our approach no replication estimates are used to estimate any parameter, all evaluations presented in this paper provide "out-of-sample" performance measures, thereby eliminating the need to split the data and perform cross-validation.

The structure of this paper is as follows: descriptive results on the data collected are presented in the following section. We then develop a novel model of effect sizes that addresses the shortcomings of the model used in previous analyses. Next, we compute forecasts for the collected data and systematically evaluate and compare them with forecasts based on the previously used model. Finally, the paper ends with a discussion of the results obtained.

2 Data

Data from all replication projects with a "one-to-one" design (i. e., one replication for one original study) that are, to our knowledge currently available, were collected. In all data sets, effect estimates were provided as correlation coefficients (r). The R code and details on data preprocessing can be found in the supplement. 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, $\hat{\theta} = \tanh^{-1}(r)$, the transformed correlation coefficients become asymptotically normally distributed with their variance only being a function of the study sample size n, i. e., $\text{Var}(\hat{\theta}) = 1/(n-3)$ (Fisher, 1921).

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

Experimental Economics Replication Project This project attempted to replicate 18 experimental economics studies published between 2011 and 2015 in two high impact economics journals (Camerer et al., 2016). For this project a *prediction market* was also conducted in order to estimate the peer beliefs about whether a replication will result in a statistically significant result. Since the estimated beliefs are also probabilistic predictions, they can be compared to the probability of a significant replication effect estimate under the statistical prediction models.

Social Sciences Replication Project This project involved 21 replications of studies on the social sciences published in the journals *Nature* and *Science* between 2010 and 2015 (Camerer et al., 2018). As in the experimental economics replication project, a prediction market to estimate peer beliefs about the replicability of the original studies was conducted and the resulting belief estimates can be used as a comparison to the statistical predictions.

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

2.1 Descriptive results

Fig 1 shows plots of the original versus the replication effect estimate, both on the correlation scale.

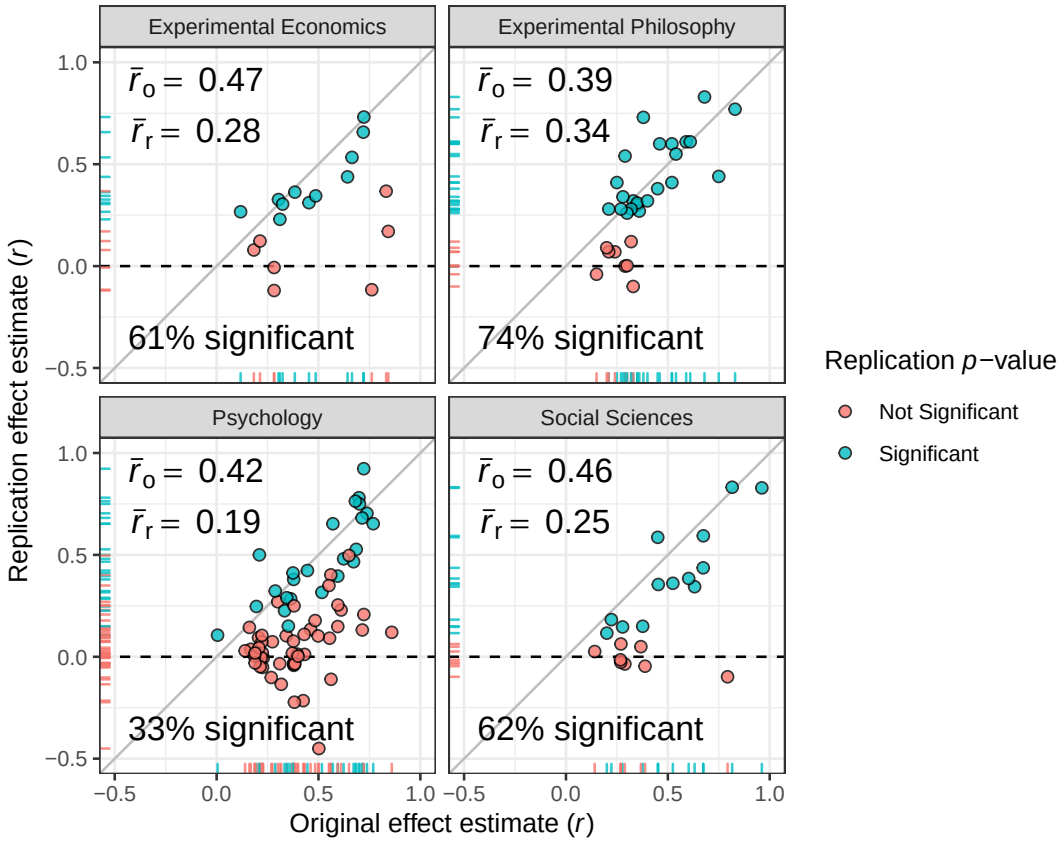

Figure 1: Original effect estimate versus replication effect estimate (on correlation scale). The color indicates whether statistical significance at the (two-sided) 0.05 level was achieved.

Most effect estimates of the replication studies are considerably smaller than those of the original studies. In particular, the mean effect estimates of the replications are roughly half as large as the mean effect estimates of the original studies. This is not the case for the philosophy project, however, 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 in the effect estimate were less likely to achieve statistical significance in the replication.

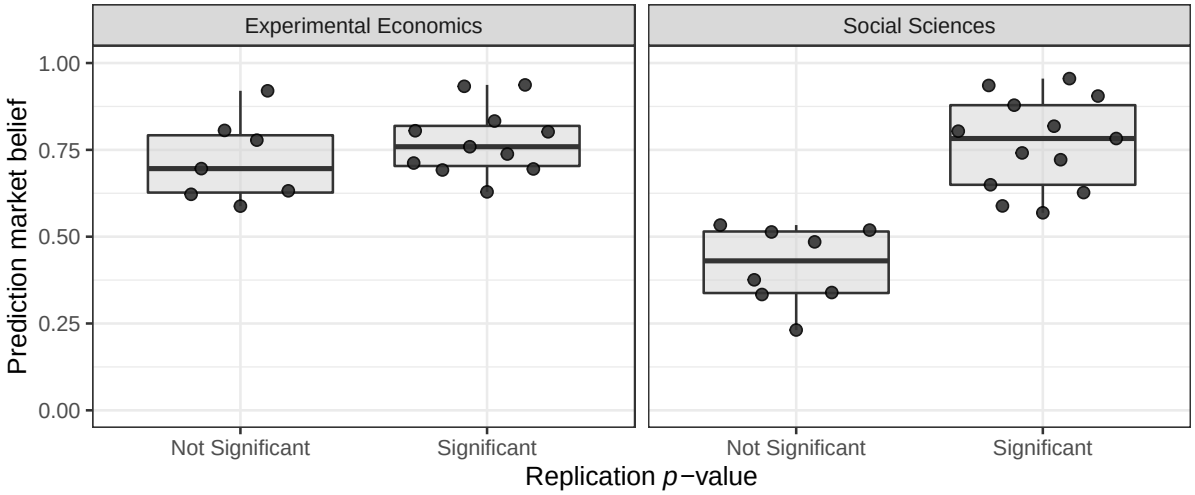

Figure 2: Statistical significance of replication effect estimate versus estimated prediction market beliefs about whether the replication will achieve statistical significance (at two-sided $\alpha = 0.05$).

Fig reffig:prediction-market-descriptive illustrates the elicited prediction market beliefs about whether the replication studies will achieve statistical significance. In the case of the economics data set, the distribution of the prediction market beliefs is very similar for significant and non-significant replications. In the social sciences project, on the other hand, the elicited beliefs separate significant and non-significant replications completely for a cut-off around 0.55.

3 Methods

To introduce some notation, denote the overall effect by θ, study-specific underlying effects by θ_o and θ_r, and their estimates by $\hat{\theta}_o$ and $\hat{\theta}_r$, with the subscript indicating whether they come from the original or the replication study. Let the corresponding standard errors be denoted by σ_o and σ_r and also let the heterogeneity variance be τ^2. Similarly, define the variance ratio as $c = \sigma_o^2/\sigma_r^2$, the relative between-study heterogeneity as $d = \tau^2/\sigma_o^2$, and also denote the corresponding test statistics by $t_o = \hat{\theta}_o/\sigma_o$ and $t_r = \hat{\theta}_r/\sigma_r$. Finally, let $\Phi(x)$ be the cumulative distribution function of the standard normal distribution evaluated at x and let z_α denote the $1-\alpha$ quantile thereof.

We propose the following Bayesian hierarchical model for the effect estimates

$$\hat{\theta}_k \,|\, \theta_k \sim \mathrm{N}(\theta_k, \sigma_k^2) \tag{1a}$$
$$\theta_k \,|\, \theta \;\sim \mathrm{N}(\theta, \tau^2) \tag{1b}$$
$$\theta \;\sim \mathrm{N}(\mu_\theta, \sigma_\theta^2) \tag{1c}$$

where $\sigma_k^2, \tau^2, \mu_\theta, \sigma_\theta^2$ are fixed and $k \in \{o, r\}$ (see Fig 3 for a graphical illustration).

After a suitable transformation a large variety of effect size measures are covered by this framework (e. g., mean differences, odds ratios, correlations). For instance, $\hat{\theta}_k = \tanh^{-1}(r_k)$ and $\sigma_k^2 = 1/(n_k - 3)$ are used in the four data sets for our analysis. The normality assumption is also common to many meta-analysis methods. Together with a fixed heterogeneity variance τ^2, it leads to analytical tractability of the predictive distributions. In this model, the case where effect estimates of the original and replication studies are not realizations of the same,

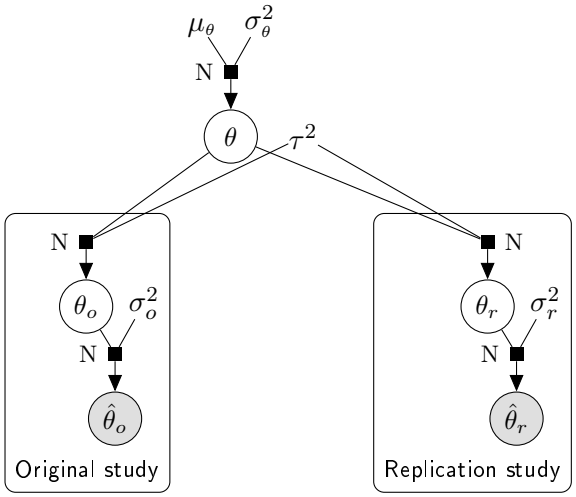

Figure 3: Hierarchical model of effect sizes in replication setting. Random variables are encircled (and grey if they are observable).

but of slightly different underlying random variables is taken into account and controlled by the heterogeneity variance τ^2. That is, for the limiting case $\tau^2 \to 0$, the study-specific underlying effects θ_o and θ_r are assumed to be the same, while for the other extreme $\tau^2 \to \infty$, θ_o and θ_r are assumed to be completely unrelated. Furthermore, the choice of the prior distribution of θ provides additional flexibility to incorporate prior knowledge about the overall effect. In the following, the predictive distributions of the replication effect estimate under two interesting prior distributions are discussed.

Flat prior If the prior distribution Eq (1c) is chosen to be flat, the posterior distribution of the overall effect θ after observing the original study effect estimate becomes $\theta \,|\, \hat{\theta}_o \sim \mathrm{N}(\hat{\theta}_o, \sigma^2_o + \tau^2)$. The posterior predictive distribution of $\hat{\theta}_r$ then turns out to be

$$\hat{\theta}_r \,|\, \hat{\theta}_o \sim \mathrm{N}\left(\hat{\theta}_o, \sigma^2_o + \sigma^2_r + 2\tau^2\right). \tag{2}$$

Under this predictive model, one implicitly assumes the effect estimation in the original study to be unbiased, since the predictive density is centered around the original effect estimate. Furthermore, the uncertainty coming from the original and replication study, as well as the uncertainty from the between-study heterogeneity is taken into account. Also note that for $\tau^2 = 0$, Eq (2) reduces to the naive model from Patil et al. (2016) used in previous analyses.

Given this predictive model, the test statistic of the replication is distributed as

$$t_r \,|\, t_o \sim \mathrm{N}\left(\sqrt{c} \cdot t_o, c + 1 + 2cd\right), \tag{3}$$

which only depends on the original test statistic t_o, the variance ratio c, and the relative heterogeneity d. From Eq (3) one can easily compute the power to obtain a statistically significant result in the replication study. Hence, this generalizes the *replication probability* (Goodman, 1992; Senn, 2002), i. e., the probability of obtaining a statistically significant finding in the same direction as in the original study, to the setting of possible between-study heterogeneity.

Sceptical prior Instead of a flat prior, one can also choose a normal prior centered around zero for Eq (1c), reflecting a more sceptical belief about the overall effect (Spiegelhalter et al., 2004, Chapter 5.5.2). Moreover, we decided to use a parametrization of the variance parameter inspired

by the g-prior (Zellner, 1986) known from the regression literature, i. e., $\theta \sim \mathrm{N}(0, g \cdot [\sigma_o^2 + \tau^2])$ with fixed $g > 0$. 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 (empirical Bayes estimation). In doing so, the empirical Bayes estimate $\hat{g} = \max\{\hat{\theta}_o^2/(\sigma_o^2 + \tau^2) - 1, 0\}$ is obtained. Fixing g to $\hat{g}$ and applying Bayes' theorem, the posterior distribution of the overall effect θ after observing the original effect estimate becomes $\theta \,|\, \hat{\theta}_o, \hat{g} \sim \mathrm{N}(s \cdot \hat{\theta}_o, s \cdot [\sigma_o^2 + \tau^2])$, with shrinkage factor

$$s = \frac{\hat{g}}{1 + \hat{g}} = \max\left\{1 - \frac{1 + d}{t_o^2}, 0\right\}. \tag{4}$$

Fig 4 shows the shrinkage factor s as a function of the relative between-study heterogeneity d and the test statistic (or the two-sided p-value) of the original study. Interestingly, for $d = 0$, Eq (4) reduces to the factor known from the theory of optimal shrinkage of regression coefficients (Copas, 1983, 1997).

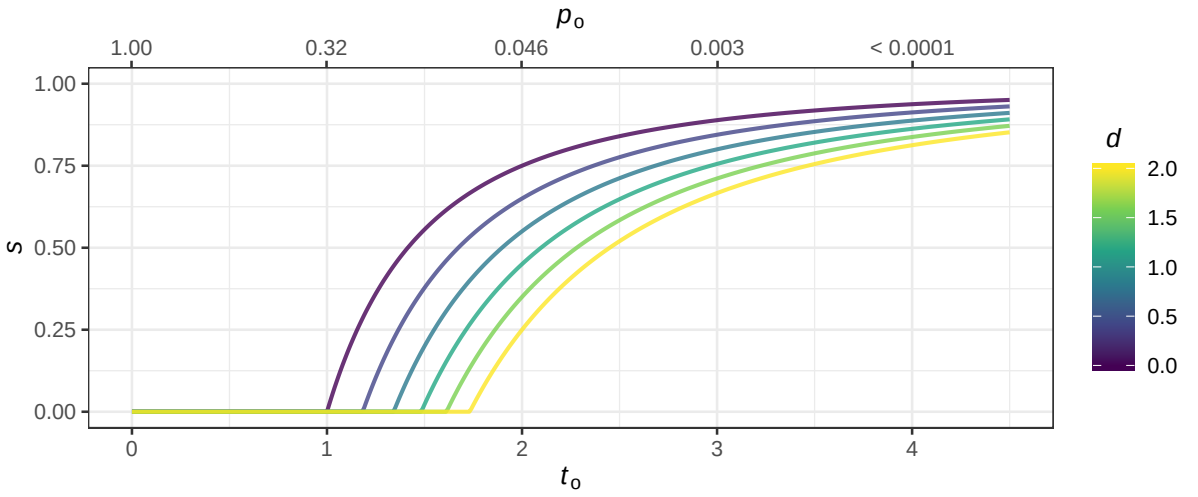

Figure 4: Shrinkage factor s as function of the test statistic t_o (bottom axis) and the two-sided p-value p_o (top axis) of the original study and the relative between-study heterogeneity $d = \tau^2/\sigma_o^2$.

The posterior predictive distribution of $\hat{\theta}_r$ under this model becomes

$$\hat{\theta}_r \,|\, \hat{\theta}_o \sim \mathrm{N}\left(s \cdot \hat{\theta}_o, s \cdot (\sigma_o^2 + \tau^2) + \sigma_r^2 + \tau^2\right). \tag{5}$$

The promise of this method is that shrinkage towards zero should improve predictive performance by counteracting the regression to the mean effect. As such, shrinkage also counteracts effect estimate inflation caused by publication bias to some extent. That is, the contribution of the original effect estimate to the predictive distribution shrinks depending on the amount of evidence in the original study (*evidence-based shrinkage*). The less convincing the result from the original study, i. e., the smaller t_o, the more shrinkage towards zero. On the other hand, shrinkage decreases for increasing evidence and in the limiting case the predictive distribution is the same as under the flat prior, i. e., $s \to 1$ for $t_o \to \infty$. Moreover, the shrinkage factor in Eq (4) is also influenced by the ratio d/t_o^2. If the test statistic is not substantially larger than the relative between-study heterogeneity, i. e., $t_o^2 \not\gg d$, heterogeneity also induces shrinkage towards zero.

Based on this predictive model, the distribution of the test statistic of the replication study

$$t_r \,|\, t_o \sim \mathrm{N}\left(s \cdot \sqrt{c} \cdot t_o, s \cdot (c + cd) + 1 + cd\right), \tag{6}$$

depends only on the relative quantities c, d, and t_o. From Eq (6), it is again straightforward to compute the power for a significant replication outcome.

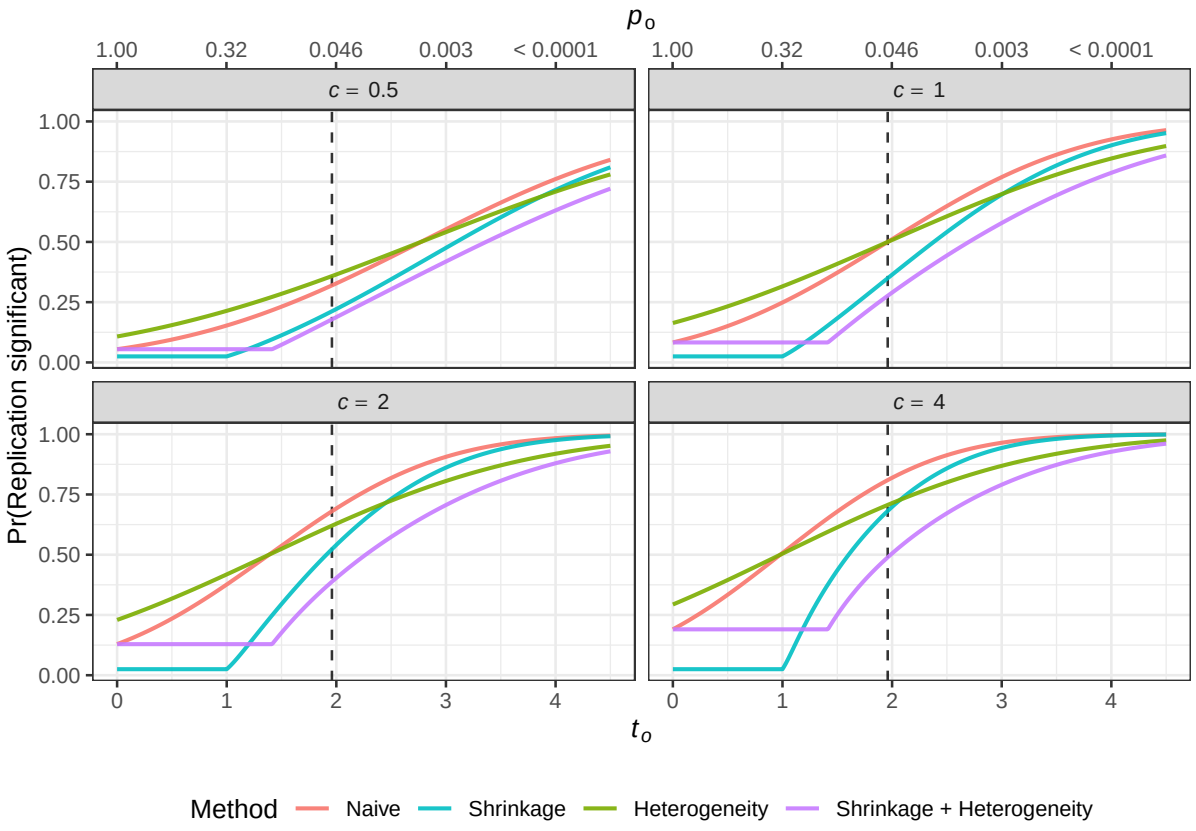

Figure 5: Probability of a significant replication outcome in the same direction as in the original study (*replication probability*) at (two-sided) $\alpha = 0.05$ level as a function of the test statistic t_o (bottom axis) and p-value p_o (top axis) of the original study and variance ratio $c = \sigma_o^2/\sigma_r^2$. The dashed line indicates $z_{0.025} \approx 1.96$. In the case of heterogeneity, $d = \tau^2/\sigma_o^2$ is set to one, otherwise to zero.

Fig 5 shows the replication probability as a function of the original test statistic t_o (or two-sided p-value p_o) and for different values of the variance ratio c. Note that the curves of the shrinkage methods stay constant until t_o reaches a point where Eq (4) starts to become larger than zero. If the original study showed a "just significant result" ($t_o \approx 1.96$) and the precision is equal in the original and the replication study ($c = 1$), the replication probability is just 0.5 when a flat prior is used. This surprising result was already noted two decades ago (Goodman, 1992), yet it has not become part of statistical literacy and many practitioners of statistics are still perplexed when they hear about it. If a sceptical prior is used, the replication probability becomes even lower. Moreover, when the precision of the replication is smaller ($c < 1$), the replication probability is also lower, whereas with increased precision ($c > 1$) the replication probability also increases. Finally, for small t_o the replication probability is higher when there is heterogeneity compared to when there is no heterogeneity, while the opposite is true for large t_o.

3.1 Specification of the heterogeneity variance

One needs to specify a value for the heterogeneity variance τ^2 to compute predictions of $\hat{\theta}_r$. However, it is not possible to estimate τ^2 using only the data from the original study, since the overall effect θ in the marginal likelihood of $\hat{\theta}_o \,|\, \theta \sim \mathrm{N}(\theta, \sigma_o^2 + \tau^2)$ is also unknown.

Ideally, a domain expert would carefully assess original study and replication protocol, and then specify how much heterogeneity can be expected for each study pair individually. This is, however, beyond the scope of this work. We instead want to compare forecasts that use a positive "default value" of τ^2 to forecasts for which τ^2 is set to zero. The goal is then to assess whether or not it makes a difference in predictive performance when heterogeneity is taken into account. Of course we will also investigate how robust our conclusions are to the choice of the default value for τ^2 by conducting a sensitivity analysis (see Section 4.3).

To determine the value of τ^2, we adapted an approach originally proposed to determine plausible values for τ^2 of heterogeneous log odds ratio effects (Spiegelhalter et al., 2004, Chapter 5.7.3, P. 168): Based on the proposed hierarchical model Eq (1), 95% of the study-specific underlying effects $\theta_k, k \in \{o, r\}$ should lie within the interval $\theta \pm z_{0.025} \cdot \tau$. We want to specify a value for τ, such that the range of this interval is not zero, but also not very large, because the whole purpose of a replication study is to replicate an original experiment as closely as possible. The comparison of the limits of this interval is, however, easier if they are transformed to the correlation scale, since θ_k are a Fisher z-transformed correlations $\theta_k = \tanh^{-1}(r_k)$ in all used data sets. We therefore looked at the difference of the transformed 97.5% to the transformed 2.5% quantile, $\delta(\tau) = \tanh(\theta_{k,97.5\%}) - \tanh(\theta_{k,2.5\%})$ as a function of the heterogeneity parameter τ for an overall effect of $\theta = 0$ (Fig 6). We then chose a value for τ that lead to a plausible value of $\delta(\tau)$.

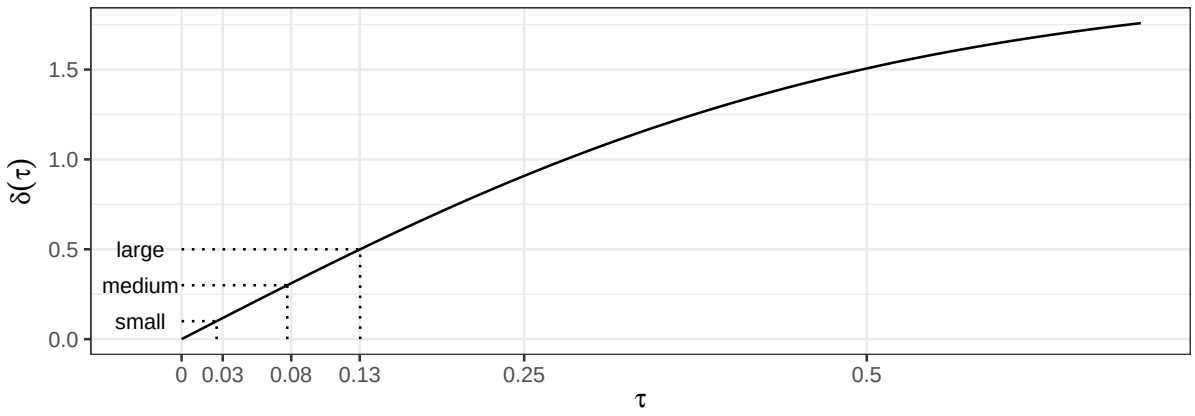

Figure 6: Difference between backtransformed quantiles $\delta(\tau) = \tanh(\theta_{k,97.5\%}) - \tanh(\theta_{k,2.5\%})$ as a function of between-study heterogeneity τ for $\theta = 0$. The values corresponding to small, medium, and large correlation effect sizes according to the classification by Cohen (Cohen, 1992) are depicted by dotted lines.

However, this raises the question of how one should classify these differences and which value should be chosen for the current setting. In the context of power analysis, there exist many classifications of effect size magnitudes, e. g., the one by Cohen (Cohen, 1992). We think this classification is appropriate since it was developed to characterize effects in psychology and other social sciences, the fields from which the data at hand are. In this classification a medium effect size should reflect an effect which is "visible to the eye" ($r = 0.3$), a small effect size should be smaller but not trivial ($r = 0.1$), and a large effect size should have the same difference to the medium effect size as the small effect size, but in the other direction ($r = 0.5$). For direct replication studies, we think it is reasonable to assume that the between-study heterogeneity should not be very large, because these kinds of studies are usually matched as closely as possible to the original studies. We therefore chose $\tau = 0.08$, such that $\delta(\tau)$, the difference between the 97.5% and 2.5% quantiles of the study-specific underlying effects, is not larger than the size of a medium effect.

An alternative approach would be to use empirical heterogeneity estimates known from the literature. We therefore also compared the chosen value to the empirical distribution of 497 between-study heterogeneity estimates of meta-analyses with correlation effect sizes in the journal *Psychological Bulletin* between 1990 and 2013 (Erp et al., 2017) (see the supplement for details). The value of 0.08 corresponds to the 34% quantile of the empirical distribution, which we think is reasonable as those estimates stem from meta-analyses of ordinary studies that are likely to be more heterogeneous than direct replication studies.

3.2 Predictive evaluation methods

A large body of methodology is available to assess the quality of probabilistic forecasts. When comparing the actual observations with their predictive distributions, one can distinguish different aspects. *Discrimination* characterizes how well a model is able to predict different observations. *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 also assess *sharpness* of the predictions, i. e., the concentration of the predictive distribution (Gneiting and Katzfuss, 2014).

Proper scoring rules are an established way to assess calibration and sharpness of probabilistic forecasts simultaneously. We therefore computed the mean logarithmic (LS), quadratic (QS), and continuous ranked probability score (CRPS) for continuous predictive distributions (Gneiting et al., 2007), and the mean (normalized) Brier score (BS) for binary predictive distributions (Schmid and Griffith, 2005). In order to specifically evaluate calibration, several methods were used: First, calibration tests based on scoring rules were conducted, i. e., *Spiegelhalter's z-test* (Spiegelhalter, 1986) for forecasts with a binary target and four calibration tests based on LS and CRPS (Held et al., 2010) for forecasts with a continuous target. All of these tests exploit the fact that under the null hypothesis of perfect calibration, the distribution of certain scores can be determined. Second, the *probability integral transform* (PIT), i. e., the value of the predictive cumulative distribution function evaluated at the actual observed value, was computed for each forecast. Under perfect calibration, the PIT values should be uniformly distributed which can be assessed visually, as well as with formal tests (Gneiting et al., 2007). Third, the *calibration slope* method was used to evaluate calibration by regressing the actual observations on their predictions, i. e., for forecasts with a binary target using logistic regression. A well calibrated prediction model should lead to a regression slope $\beta \approx 1$, whereas $\beta > 1$ and $\beta < 1$ indicate miscalibration (Cox, 1958). Finally, to assess the discriminative quality of the forecasts with a binary target, the *area under the curve* (AUC) was computed (Steyerberg, 2009, Chapter 15.2.3).

3.3 Software

All analyses were performed in the R programming language (R Core Team, 2019). The full code to reproduce analyses, plots, and tables is provided in the supplement. Methods to compute prediction intervals and to conduct sample size calculations (see the supplement for details), as well as the four data sets are provided in the R package `ReplicationSuccess` which is available at `https://r-forge.r-project.org/projects/replication/`.

4 Results

In this section, predictive evaluations of four different forecasting methods applied to the data sets are shown: the method with the flat prior and $\tau = 0$, corresponding to the previously used method from Patil et al. (2016) (denoted by N for *naive*), the method with the flat prior and $\tau = 0.08$ (denoted by H for *heterogeneity*), the method with the sceptical prior and $\tau = 0$

(denoted by S for *shrinkage*), and the method with the sceptical prior and $\tau = 0.08$ (denoted by SH for *shrinkage and heterogeneity*).

4.1 Forecasts of effect estimates

In the following, evaluations of forecasts of the replication effect estimates are shown.

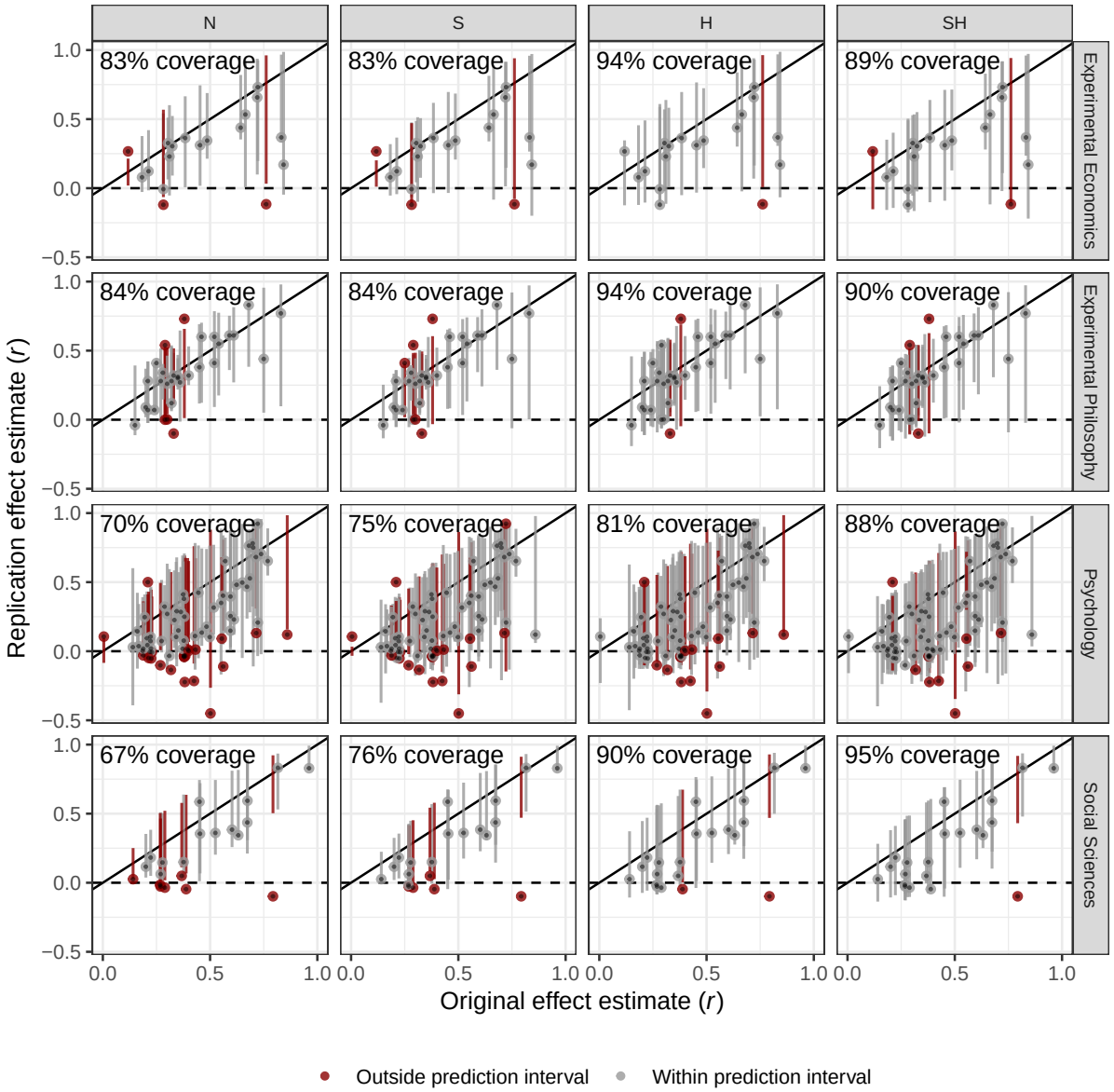

Figure 7: Original and replication effect estimates with 95% prediction intervals of the replication effect estimates (vertical lines). Forecasting methods are abbreviated by N for *naive*, S for *shrinkage*, H for *heterogeneity*, SH for *shrinkage and heterogeneity*.

Prediction intervals Fig 7 shows plots of the original versus the replication effect estimates. In addition, the corresponding 95% prediction interval is vertically shown around each study pair. Comparing the different methods across projects, the S method shows similar coverage as the N method in the economics and philosophy data sets (83 - 84%), whereas in the psychology and social sciences data sets the S method (75 - 76%) shows a higher coverage compared to the

N method (67 - 70%). As expected, when heterogeneity is taken into account, the prediction intervals become wider and the coverage improves considerably in all cases. In the philosophy and economics projects the highest coverage is achieved for the forecasts from the H method (94%), while in the psychology and social sciences projects the highest coverage is achieved for the SH forecasts (88 - 95%). Moreover, in all but the psychology data set, the best method is able to achieve nominal coverage, whereas in the psychology data set the best method achieves slightly less. These improvements suggest improved calibration of the forecasts which take heterogeneity into account (and shrinkage in the case of the social sciences and psychology data sets). Finally, in the psychology and social sciences projects, the replication effect estimates that are not covered by their prediction intervals tend to be smaller than the lower limits of the intervals. In the economics and philosophy projects, on the other hand, the non-coverage appears to be more symmetric.

Scores Table 1 shows the mean quadratic score (QS), mean logarithmic score (LS), and the mean continuous ranked probability score (CRPS) for each combination of data set and forecasting method. The SH method achieved the lowest mean score for each score type and in all projects, suggesting that this method performs the best among the four methods. The N method, on the other hand, usually showed the highest mean score across all score types, indicating that this method performs worse compared to the other methods. We also tested for deviation from equal predictive performance using paired tests and in most cases there is evidence that the difference between the scores of the SH forecasts and the scores of the other forecasts is substantial (see the supplement for details).

Table 1: Mean quadratic score (QS), mean logarithmic score (LS), mean continuous ranked probability score (CRPS), and harmonic mean of p-values from four score-based calibration tests ($\mathring{p}$). Forecasting methods are abbreviated by N for *naive*, S for *shrinkage*, H for *heterogeneity*, SH for *shrinkage and heterogeneity*.

		Score Type			
Project	Method	QS	LS	CRPS	$\mathring{p}$
Experimental Economics	N	−0.83	0.34	0.21	0.013
$n = 18$	S	−1.17	0.17	0.17	0.056
	H	−1.14	0.18	0.21	0.24
	SH	−1.32	0.02	0.17	0.79
Experimental Philosophy	N	−1.33	−0.05	0.12	0.0005
$n = 31$	S	−1.46	−0.06	0.12	0.0002
	H	−1.51	−0.18	0.12	0.81
	SH	−1.67	−0.20	0.11	0.66
Psychology	N	−0.07	0.87	0.22	< 0.0001
$n = 73$	S	−0.15	0.86	0.19	< 0.0001
	H	−0.55	0.51	0.22	< 0.0001
	SH	−0.85	0.27	0.18	< 0.0001
Social Sciences	N	−0.17	0.85	0.22	< 0.0001
$n = 21$	S	−0.58	0.54	0.19	< 0.0001
	H	−0.67	0.55	0.21	< 0.0001
	SH	−1.17	0.25	0.18	0.01

Calibration tests A total of four score-based calibration tests have been performed. These tests exploit the fact that for normal predictions under the null hypothesis of perfect calibration,

the first two moments of the distribution of the mean LS and the mean CRPS can be derived and appropriate unconditional calibration tests can be constructed. Moreover, the functional relationship between the two moments can be used to define a regression model in which the individual scores are regressed on their (suitably transformed) predictive variances leading to another procedure to test for miscalibration (Held et al., 2010). A theoretically well-founded way to summarize the p-values of these tests is to use their harmonic mean $\mathring{p}$ (Good, 1958; Wilson, 2019; Held, 2019), which is also shown in Table 1 (see the supplement for non-summarized results).

Taken together, there is strong evidence for miscalibration of all forecasts in the psychology and social sciences projects. In the economics project, on the other hand, there is no evidence for miscalibration of the H and SH forecasts and weak evidence for miscalibration of the other forecasts. Finally, in the philosophy project there is strong evidence for miscalibration of the N and S forecasts and no evidence for miscalibration of the H and SH forecasts.

PIT histograms Fig 8 shows histograms of the PIT values of the four forecasting methods along with p-values from Kolmogorov-Smirnov tests for uniformity.

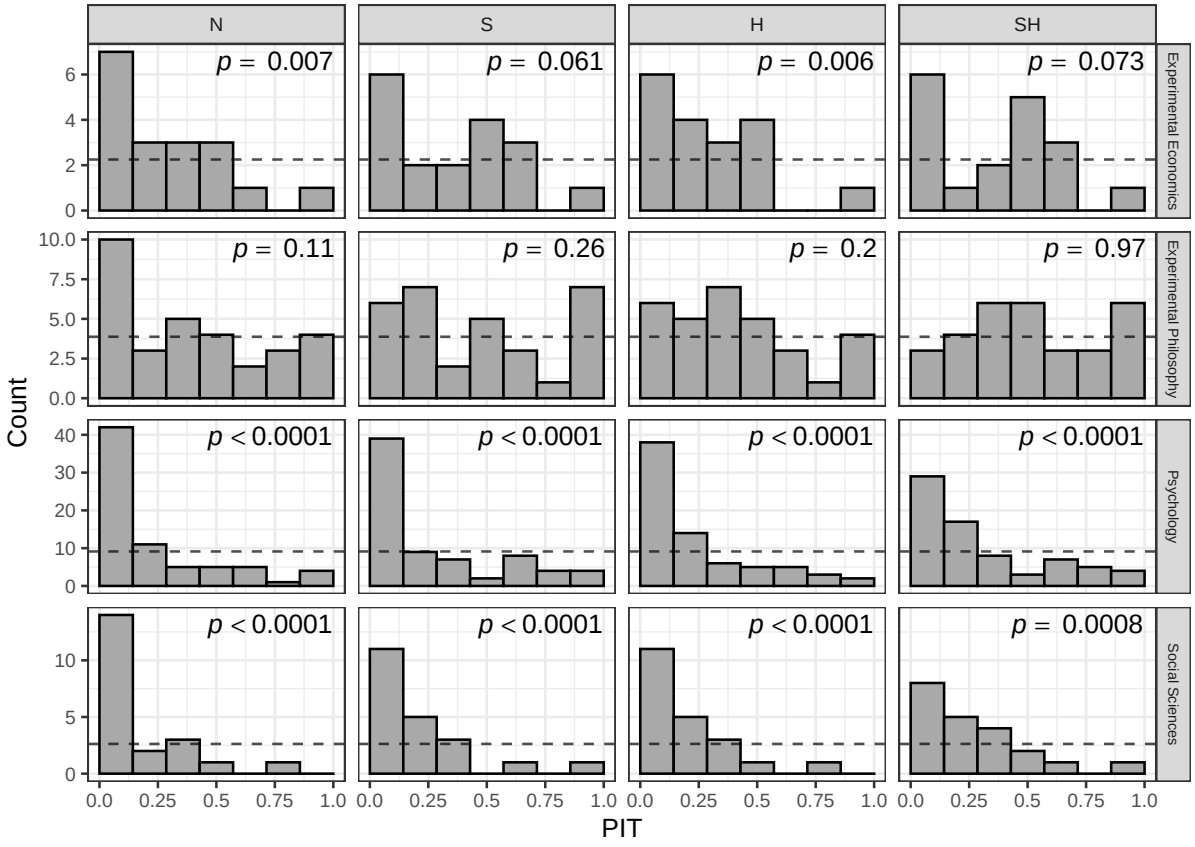

Figure 8: Histograms of PIT values with p-values from Kolmogorov-Smirnov test for uniformity. Dashed lines indicate number of counts within bins expected under uniformity. Forecasting methods are abbreviated by N for *naive*, S for *shrinkage*, H for *heterogeneity*, SH for *shrinkage and heterogeneity*.

In some of the histograms in the social sciences and economics projects there are bins with zero observations, however, these projects also have the smallest sample sizes. In the psychology and social sciences data sets, the N method shows extreme bumps in the lower range of the PIT values, while the histograms of the H, S, and SH methods look flatter, suggesting less

miscalibration. In the economics data set, the PIT histograms also show bumps in the lower range, but to a much lower degree than in the psychology and social sciences data sets. Finally, in the philosophy data set the histograms look acceptable for all methods, suggesting no severe miscalibration.

4.2 Forecasts of statistical significance

Since statistical significance of the replication study is one of the most commonly used criteria for replication success, in the following section the probability of significance under the investigated predictive distributions will be evaluated using methods suited for probabilistic forecasts of binary target variables. Moreover, in the social sciences and experimental economics projects these forecasts can also be compared to forecasts from the prediction markets. The significance threshold $\alpha = 0.05$ for a two-sided p-value was used in all cases. Most of the evaluations were also conducted for smaller α thresholds and are reported in the supplement.

Expected number of statistically significant replication studies By summing up all probabilities for significance under each method within one project, the expected number of statistically significant replication outcomes is obtained and can be compared to the observed number, e. g., with a χ^2-goodness-of-fit test (shown in Table 2).

Table 2: Observed and expected number of statistically significant replication studies along with p-value from χ^2-goodness-of-fit test. Forecasting methods are abbreviated by N for *naive*, S for *shrinkage*, H for *heterogeneity*, SH for *shrinkage and heterogeneity*, PM for *prediction market*.

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

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. The overestimation is the smallest for the SH method and the largest for the N method across all data sets.

In the economics and philosophy projects there is no evidence of a difference between expected and observed under the S and SH method, whereas there is weak to moderate evidence of a difference for the N and H methods. In the social sciences and psychology projects, on the

other hand, there is strong evidence for a difference between the expected and the observed number of significant replications for all methods, suggesting miscalibration of these forecasts. Furthermore, the expected numbers under the prediction market (PM) method in the economics and social sciences projects do not differ substantially from what was actually observed, providing no evidence for miscalibration of these forecasts.

Brier scores In Table 3 the mean (normalized) Brier scores are shown for each combination of data set and forecasting method. The mean normalized Brier score (Schmid and Griffith, 2005) is shown because it enables the comparison of models across data sets in which the proportion of significant replications differs (e. g., in the psychology data set the proportion is much lower than in the others). It is computed by $\text{BS}_\text{n} = (\text{BS}_0 - \text{BS})/\text{BS}_0$ where BS_0 is the baseline Brier score assuming that all replication studies are given an estimated probability of significance equal to the proportion of significant replications. Hence, BS_n is positive if the predictive performance of the model is better than the baseline prediction.

Table 3: Mean Brier score (BS), mean normalized Brier score (BS norm), and test-statistic with p-value from Spiegelhalter's z-test. Forecasting methods are abbreviated by N for *naive*, S for *shrinkage*, H for *heterogeneity*, SH for *shrinkage and heterogeneity*, PM for *prediction market*.

Project	Method	BS	BS norm	z	p-value
Experimental Economics	N	0.271	−0.139	2.49	0.013
$n = 18$	S	0.226	0.048	1.15	0.25
	H	0.262	−0.104	2.03	0.042
	SH	0.227	0.046	0.79	0.43
	PM	0.243	−0.021	1.45	0.15
Experimental Philosophy	N	0.193	−0.007	3.31	0.0009
$n = 31$	S	0.173	0.097	2.07	0.039
	H	0.170	0.110	1.60	0.11
	SH	0.148	0.229	0.21	0.83
Psychology	N	0.394	−0.784	10.00	< 0.0001
$n = 73$	S	0.335	−0.518	7.81	< 0.0001
	H	0.363	−0.644	8.50	< 0.0001
	SH	0.289	−0.308	5.17	< 0.0001
Social Sciences	N	0.346	−0.468	7.20	< 0.0001
$n = 21$	S	0.324	−0.374	5.46	< 0.0001
	H	0.310	−0.316	4.89	< 0.0001
	SH	0.272	−0.155	3.42	0.0006
	PM	0.114	0.519	− 2.02	0.044

In the social sciences and psychology projects the predictive performance is poor for all statistical methods. Namely, all mean Brier scores are larger than 0.25, a score that can be obtained by simply using 0.5 as estimated probability every time and additionally all mean normalized Brier scores are negative. In the economics project, the S and SH methods achieve a positive mean normalized Brier score, while it is negative for the N and H methods. Finally, the forecasts in the philosophy project show the best performance, i. e., all methods except the N method achieve a positive mean normalized Brier score with the SH method showing the largest value. Moreover, the PM forecasts show a normalized Brier score of about zero in the economics projects, which is comparable to the statistical methods, whereas in the social sciences project, the performance is remarkably good, far better than all statistical forecasts in this project.

Table 3 also displays the results of Spiegelhalter's z-test. In the psychology and social sciences

data sets the test provides strong evidence for miscalibration of all statistical forecasts, but only weak evidence for miscalibration of the PM forecasts in the social sciences data set. In the economics data set, on the other hand, there is no evidence for miscalibration of the S, SH and the PM forecasts and weak evidence for miscalibration of the N and H forecasts. Finally, in the philosophy data set there is moderate evidence for miscalibration of the N and S forecasts, but no evidence for miscalibration of the H and SH forecasts.

Calibration slope Fig 9A shows the calibration slopes obtained by logistic regression of the outcome whether the replication achieved statistical significance on the logit transformed estimated probabilities. In all but the psychology project the confidence intervals are very wide due to the small sample size. Also note that it was not possible to obtain the calibration slope for the PM method in the social science project because of complete separation. In the psychology and social sciences projects, the slopes of all methods are considerably below the nominal value of one suggesting miscalibration. However, the H and SH methods show higher values than the methods that do not take heterogeneity into account, indicating improvements in calibration. In the economics and philosophy projects, the slopes of all methods are closer to one and all confidence intervals include one, suggesting no miscalibration.

Area under the curve Fig 9B shows the area under the curve (AUC) for each combination of data set and forecasting method. The 95% Wald type confidence intervals were computed on the logit scale and then backtransformed.

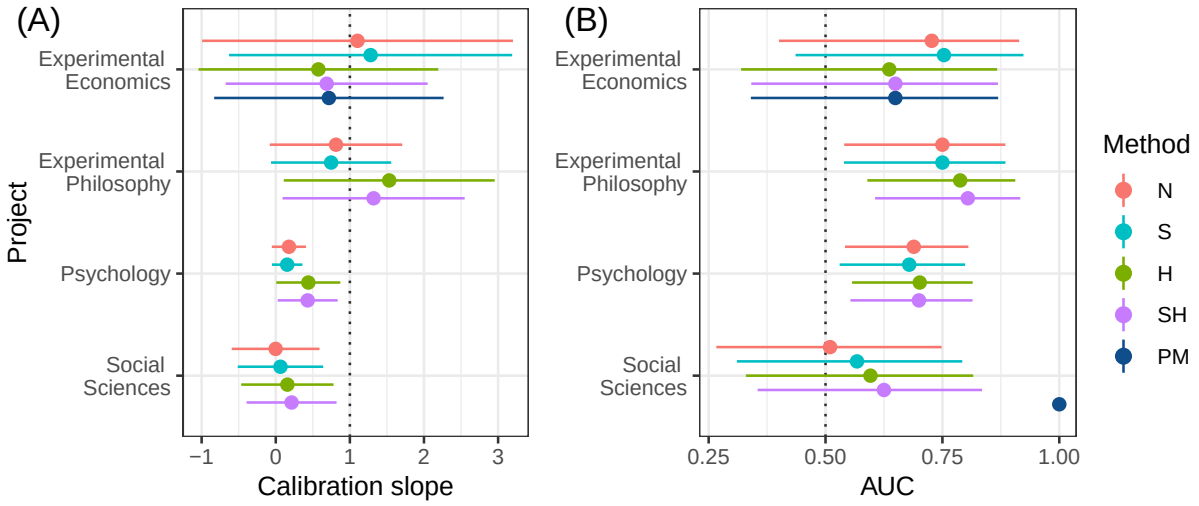

Figure 9: Calibration slope and area under the curve (AUC) with 95% confidence interval. Forecasting methods are abbreviated by N for *naive*, S for *shrinkage*, H for *heterogeneity*, SH for *shrinkage and heterogeneity*, PM for *prediction market*.

Note that in the social sciences project for the PM forecasts, an AUC of one (without confidence interval) was obtained because the forecasts were able to completely separate non-significant and significant replications. The statistical forecasts in the social sciences project, 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 projects the H and SH methods show the highest AUCs. The former are around 0.8, while the latter are about 0.7, indicating reasonable discriminatory power of all forecasts. Finally, in the economics data set the N and S methods achieve the highest AUCs with values of around 0.75, but with very wide confidence intervals which all include 0.5.

4.3 Sensitivity analysis of heterogeneity variance choice

For the H and SH methods the heterogeneity parameter τ was set to a value of 0.08 as described earlier. We performed a sensitivity analysis to investigate how much the results change when other values are selected. The change in predictive performance was investigated using the mean QS, mean LS, and mean CRPS, as they are good summary measures for calibration and sharpness of a predictive distribution. Furthermore, optimizing the mean score has been proposed as a general method for parameter estimation, which also includes maximum likelihood estimation (i. e., optimum score estimation based on the LS) (Gneiting and Raftery, 2007, Section 9).

Fig 10 shows the the mean scores for each project as a function of the heterogeneity τ.

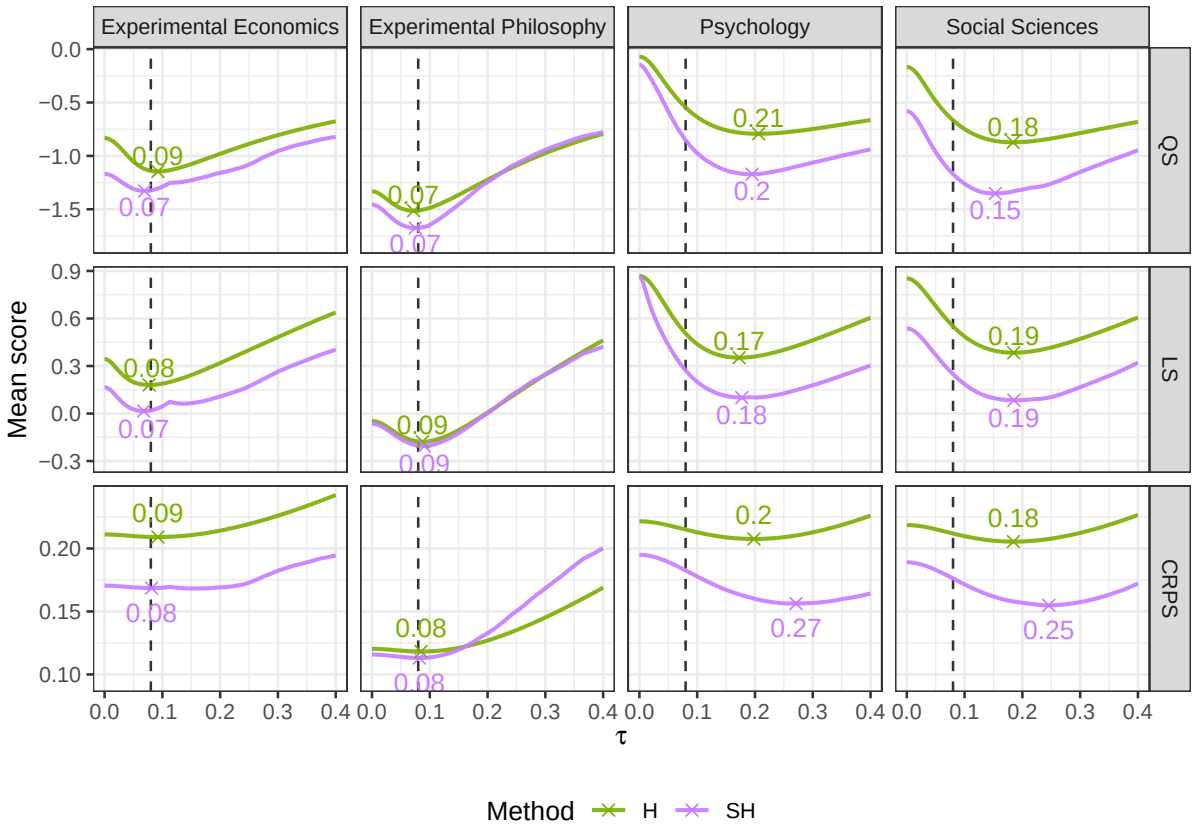

Figure 10: Mean scores as a function of τ for each score type and project. The dashed line indicates the chosen value of 0.08. Minima are indicated by a cross. Forecasting methods are abbreviated by H for *heterogeneity* and SH for *shrinkage and heterogeneity*.

In general, many of the mean score functions are rather flat, suggesting large uncertainty about the τ parameter. However, the chosen value of 0.08 seems plausible for the economics and philosophy projects, as it is close to the minima of all mean score functions. The values of τ, which minimize the mean score functions in the social sciences and psychology projects, on the other hand, are substantially larger than 0.08. The SH model shows smaller mean scores than the H model over the entire range of τ in all but the philosophy data set, where both models show comparable mean scores. This suggests that evidence-based shrinkage leads to a better (or at least equal) predictive performance across all data sets and that the comparison of the methods is not severely influenced by the choice of τ.

5 Discussion

This paper 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. In all models we assumed that after a suitable transformation an effect can be modelled by a normally distributed random variable. Furthermore, we either assumed that in the original study the effect was estimated in an unbiased way (*naive model*), or we shrunk the effect towards zero based on the evidence in the original study (*shrinkage model*). In a Bayesian framework, the former arises when choosing a flat prior distribution for the effect, while the latter arises by choosing a zero-mean normal prior and estimating the prior variance 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 incorporated by a hierarchical model structure of the effect sizes.

Replication has been investigated from a predictive point of view before; Bayarri and Mayoral (2002) used a similar hierarchical model but chose a full Bayesian approach with priors put also on the variance parameters. For the overall effect, on the other hand, they chose a flat prior, which leads to a predictive distributions without shrinkage towards zero. Patil et al. (2016) used a simpler model which was derived in a non-Bayesian framework, but corresponds to our naive model. This model was then used to obtain forecasts of replication effect estimates using the data set from the *Reproducibility Project: Psychology* (Open Science Collaboration, 2015) and also 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 forecasts. For this reason, we computed and evaluated the predictive distributions under the four different models for the three aforementioned data sets and additionally for the data from the *Experimental Philosophy Replicability Project* (Cova et al., 2018).

Predictive evaluation By taking into account between-study heterogeneity, evidence-based shrinkage, or both, calibration and sharpness have improved compared to the naive method. Forecasts 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 have been larger in the social sciences and psychology and smaller in the economics and philosophy projects. However, in the psychology and social sciences projects, 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 projects.

Furthermore, in the social sciences and economics data sets, the forecasts could be compared to forecasts from 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.

It seems likely that in many of the investigated fields there is between-study heterogeneity present, as the models that take heterogeneity into account always performed the same or better than their counterparts which do not take heterogeneity into account. This is not surprising, as many of the replications used for example samples from different populations or different materials than those in the original studies (Gilbert et al., 2016). Evidence-based shrinkage also improved predictive performance considerably in most cases, indicating that shrinkage is necessary to counteract regression to the mean. Moreover, this could suggest that the effect estimates

from the original studies were to some degree inflated or even false positives, e. g., because of publication bias or the use of questionable research practices.

Differences between replication projects The predictive performance differed between the replication projects. 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., calibration tests. That is, the psychology project consists of many more study pairs, which leads to higher power to detect miscalibration in this project compared to the other projects.

Another explanation might be that differences in the study selection process of the replication projects lead to the observed differences. For instance, the original studies in the social sciences project were selected from the journals *Nature* and *Science*, which are known to mainly promote novel and exciting research, while in the philosophy, economics, and psychology projects they were selected from standard journals. 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.

It may also be the case that the degree of inflation of original effect estimates varies between the different fields and that this leads to the observed differences. In particular, in the economics, social sciences, and psychology projects, the predictive performance was more substantially improved through evidence-based shrinkage than in the philosophy project, although the amount of shrinkage was roughly the same in all projects (see the supplement for details). One possible explanation might be that experimental philosophy is less susceptible to publication bias, as it is a much younger field where there is high acceptance for negative or null results (Cova et al., 2018). However, it may also be that in the early days of a field more obvious and more robust effects are investigated, which could explain the higher replicability of experimental philosophy findings.

Conclusions The attempt to forecast the results of replication studies brought new insights. 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 previously used naive model (Patil et al., 2016), which assumes that the effect estimates of the original study are not inflated and that there is no between-study heterogeneity.

These results have various implications: First, state-of-the-art methods for assessing discrimination, calibration, and sharpness should be used to evaluate probabilistic forecasts of replication outcomes. This allows to make more precise statements about the quality of the forecasts compared to the ad-hoc methods which were used so far (Dreber et al., 2015; Patil et al., 2016; Camerer et al., 2016, 2018; Forsell et al., 2019). Second, researchers should be aware of the fact that original and replication effect estimates may show some degree of heterogeneity, although the study designs are as closely matched as possible. Finally, for the design of a new replication study, the developed model can also be used to determine the sample size required to obtain a significant replication result for a specified power. Our method pro-

vides a more principled approach compared to just shrinking the target effect size ad hoc by an arbitrary amount as was done in the planning of previous replication studies. Software for doing this as well as the four data sets are available in the R package `ReplicationSuccess` (`https://cran.r-project.org/package=ReplicationSuccess`).

However, in the analysis of replication studies it may not be a good idea to reduce replication success solely to whether or not a replication study achieves statistical significance. One reason for this is that replication studies are often not sufficiently powered (Anderson and Maxwell, 2017), so from a predictive point of view it is then not unlikely that non-significance will occur, even if the underlying effect is not zero. Another problem is that significance alone does not take into account effect size, i. e., significance can still be achieved by increasing the sample size of the replication study, even if there is substantial shrinkage of the replication estimate. We recommend instead to 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, 2020) are promising approaches to replace statistical significance as the main criterion for replication success.

Our results also offer interesting insights about the predictability of replication outcomes in four different fields. However, they should not be interpreted to mean 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 (see the discussion in the section "Differences between replication projects"). The complexity underlying any replication project is enormous, we should applaud all the researchers involved for investing their limited resources into these 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 paper also has some limitations: In all models, the simplifying assumption of normally distributed likelihood and prior has been made, which can be questionable for smaller sample sizes. Moreover, a pragmatic Bayesian approach was chosen, i. e., no prior was put on the heterogeneity variance τ^2 and the variance hyperparameter of θ was specified with empirical Bayes. We recognize that a full Bayesian treatment, such as in Bayarri and Mayoral (2002), could reflect the uncertainty more accurately. However, our strategy leads to analytical tractability of the predictive distribution. This facilitates interpretability and allows to easily study limiting cases, which would be harder for a full Bayes approach where numerical or stochastic approximation methods are required. Moreover, it is well known that shrinkage is necessary for the prediction of new observations. The empirical Bayes shrinkage factor has proven to be optimal in very general settings (Copas, 1983, 1997) and is for example also employed in clinical prediction models (Steyerberg, 2009, Chapter 13.2). Furthermore, the data sets used all come from relatively similar fields of academic science. It would also be of interest to perform the same analysis on data from the life sciences, as well as for non-academic research. Finally, only data from replication projects with "one-to-one" design were considered. It would also be interesting to conduct similar analyses for data from replication projects which use "many-to-one" replication designs, such as the "Many Labs" project (Klein et al., 2014; Ebersole et al., 2016; Klein et al., 2018), especially for the assessment of heterogeneity.

Acknowledgments

We thank Kelly Reeve for helpful comments on the draft of the manuscript.

References

Altmejd, A., Dreber, A., Forsell, E., Huber, J., Imai, T., Johannesson, M., Kirchler, M., Nave, G., and Camerer, C. (2019). Predicting the replicability of social science lab experiments.

PLOS ONE, 14(12):e0225826. doi:10.1371/journal.pone.0225826.

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

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

Begley, C. G. and Ioannidis, J. P. (2015). Reproducibility in science. *Circulation Research*, 116(1):116–126. doi:10.1161/circresaha.114.303819.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351:1433–1436. doi:10.1126/science.aaf0918.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikenstein, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E., and Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior*, 2:637–644. doi:10.1038/s41562-018-0399-z.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1):155–159.

Copas, J. B. (1983). Regression, prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society*, 45:311–354. doi:10.1111/j.2517-6161.1983.tb01258.x.

Copas, J. B. (1997). Using regression models for prediction: shrinkage and regression to the mean. *Statistical Methods in Medical Research*, 6:167–183. doi:10.1177/096228029700600206.

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. *Review of Philosophy and Psychology*. doi:10.1007/s13164-018-0400-9.

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, 45(3/4):562–565. doi:10.2307/2333203.

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

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(7):e66844. doi:10.1371/journal.pone.0066844.

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. doi:10.1016/j.jesp.2015.10.012.

Erp, S. V., Verhagen, J., Grasman, R. P. P. P., and Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in *psychological bulletin* from 1990-2013. *Journal of Open Psychology Data*, 5(1):4. doi:10.5334/jopd.33.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLOS ONE*, 4(5):e5738. doi:10.1371/journal.pone.0005738.

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

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., and Dreber, A. (2019). Predicting replication outcomes in the many labs 2 study. *Journal of Economic Psychology*, 75:102117. doi:10.1016/j.joep.2018.10.009.

Gilbert, D. T., King, G., Pettigrew, S., and Wilson, T. D. (2016). Comment on "estimating the reproducibility of psychological science". *Science*, 351(6277):1037–1040. doi:10.1126/science.aad7243.

Gneiting, T. (2008). Editorial: Probabilistic forecasting. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2):319–321. doi:10.1111/j.1467-985x.2007.00522.x.

Gneiting, T., Balabdaoui, F., and Raftery, E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:243–268. doi:10.1111/j.1467-9868.2007.00587.x.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151. doi:10.1146/annurev-statistics-062713-085831.

Gneiting, T. and Raftery, E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the Amerian Statistical Association*, 102:359–377. doi:10.1198/016214506000001437.

Good, I. J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association*, 53(284):799–813.

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

Held, L. (2019). On the Bayesian interpretation of the harmonic mean p-value. *PNAS*, 116(13):5855–5856. doi:10.1073/pnas.1900671116.

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

Held, L., Rufibach, K., and Balabdaoui, F. (2010). A score regression approach to assess calibration of continuous probabilistic predictions. *Biometrics*, 66:1295–1305. doi:10.1111/j.1541-0420.2010.01406.x.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8):e124. doi:10.1371/journal.pmed.0020124.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5):524–532. doi:10.1177/0956797611430953.

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

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(20):2781–2793. doi:10.1002/sim.6525.

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. doi:10.1027/1864-9335/a000178.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, J., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., 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(4):443–490. doi:10.1177/2515245918810225.

Ly, A., Etz, A., Marsman, M., and Wagenmakers, E.-J. (2018). Replication Bayes factors from evidence updating. *Behavior Research Methods*, 51(6):2498–2508. doi:10.3758/s13428-018-1092-x.

McShane, B. B., Tackett, J. L., Böckenholt, U., and Gelman, A. (2019). Large-scale replication projects in contemporary psychological research. *The American Statistician*, 73(sup1):99–105. doi:10.1080/00031305.2018.1505655.

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

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

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Schmid, C. H. and Griffith, J. L. (2005). Multivariate classification rules: Calibration and discrimination. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*, volume 5, pages 3491–3497. Wiley, 2nd edition.

Senn, S. (2002). Letter to the editor: A comment on replication, p-values and evidence by S. N. Goodman, Statistics in Medicine 1992; 11:875–879. *Statistics in Medicine*, 21(16):2437–2444. doi:10.1002/sim.1072.

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

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

Steyerberg, E. (2009). *Clinical Prediction Models*. Springer-Verlag New York. doi:10.1007/978-0-387-77244-8.

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *PNAS*, 116(4):1195–1200. doi:10.1073/pnas.1814092116.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Goel, P. and Zellner, A., editors, *Bayesian Inference and Decision techniques: Essays in Honor of Bruno de Finetti*, volume 6 of *Studies in Bayesian Econometrics and Statistics*, pages 233–243. Amsterdam: North-Holland.

Computational details

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

## 2022-12-21 10:27:50 Etc/UTC

sessionInfo()

## R version 3.6.2 (2019-12-12)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Debian GNU/Linux 10 (buster)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib/x86_64-linux-gnu/libopenblasp-r0.3.5.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=C
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] biostatUZH_1.8.0 tables_0.8.8     Hmisc_4.3-1      Formula_1.2-3
##  [5] survival_3.1-8   lattice_0.20-38  ggpubr_0.2.5     magrittr_1.5
##  [9] ggbeeswarm_0.6.0 ggplot2_3.2.1    tidyr_1.0.2      dplyr_0.8.4
## [13] knitr_1.28
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3          png_0.1-7           assertthat_0.2.1
##  [4] digest_0.6.25       plyr_1.8.5          R6_2.4.1
##  [7] backports_1.1.5     acepack_1.4.1       evaluate_0.14
## [10] highr_0.8           pillar_1.4.3        rlang_0.4.4
## [13] lazyeval_0.2.2      rstudioapi_0.11     data.table_1.12.8
## [16] rpart_4.1-15        Matrix_1.2-18       checkmate_2.0.0
## [19] labeling_0.3        splines_3.6.2       stringr_1.4.0
## [22] foreign_0.8-72      htmlwidgets_1.5.1   munsell_0.5.0
## [25] compiler_3.6.2      vipor_0.4.5         xfun_0.12
## [28] pkgconfig_2.0.3     base64enc_0.1-3     htmltools_0.4.0
## [31] nnet_7.3-12         tidyselect_1.0.0    tibble_2.1.3
## [34] gridExtra_2.3       htmlTable_1.13.3    fansi_0.4.1
## [37] viridisLite_0.3.0   crayon_1.3.4        withr_2.1.2
## [40] grid_3.6.2          gtable_0.3.0        lifecycle_0.1.0
## [43] scales_1.1.0        cli_2.0.1           stringi_1.4.6
## [46] farver_2.0.3        ggsignif_0.6.0      reshape2_1.4.3
## [49] latticeExtra_0.6-29 ellipsis_0.3.0      vctrs_0.2.3
```

```
## [52] cowplot_1.0.0       boot_1.3-23         RColorBrewer_1.1-2
## [55] tools_3.6.2         glue_1.3.1          beeswarm_0.2.3
## [58] purrr_0.3.3         jpeg_0.1-8.1        colorspace_1.4-1
## [61] cluster_2.1.0
```