Predicting the Replicability of Social Science Lab Experiments*

Adam Altmejd[1,2], Anna Dreber[1,3], Eskil Forsell[1], Teck Ho[6], Juergen Huber[3], Taisuke Imai[5], Magnus Johannesson[1], Michael Kirchler[3], Gideon Nave[4], and Colin Camerer[7]

[1]Stockholm School of Economics
[2]SOFI, Stockholm University
[3]Universität Innsbruck
[4]The Wharton School, University of Pennsylvania
[5]LMU Munich
[6]National University of Singapore
[7]California Institute of Technology

January 16, 2019

Abstract

We measure how accurately replication of experimental results can be predicted by a black-box statistical model. With data from four large-scale replication projects in experimental psychology and economics, and techniques from machine learning, we train a predictive model and study which variables drive predictable replication.

The model predicts binary replication with a cross validated accuracy rate of 70% (AUC of 0.79) and relative effect size with a Spearman ρ of 0.38. The accuracy level is similar to the market-aggregated beliefs of peer scientists (Camerer et al., 2016; Dreber et al., 2015). The predictive power is validated in a pre-registered out of sample test of the outcome of Camerer et al. (2018b), where 71% (AUC of 0.73) of replications are predicted correctly and effect size correlations amount to $\rho = 0.25$.

Basic features such as the sample and effect sizes in original papers, and whether reported effects are single-variable main effects or two-variable interactions, are predictive of successful replication. The models presented in this paper are simple tools to produce cheap, prognostic replicability metrics. These models could be useful in institutionalizing the process of evaluation of new findings and guiding resources to those direct replications that are likely to be most informative.

*Generous support was provided by the Alfred P. Sloan Foundation (grant G-201513929), the Chen Center for Social and Decision Neuroscience (CFC), the Austrian Science Fund FWF (SFB F63, P29362-G27), Austrian Science Foundation FWF (P 29362-G27), The Jan Wallander and Tom Hedelius Foundation (P2015-0001:1), The Knut and Alice Wallenberg Foundation (Wallenberg Academy Fellows grant to A.D.), The Swedish Foundation for Humanities and Social Sciences (NHS14-1719), Handelsbanken (P14-0214 and P13-0156) and Forte (2016-07099). We thank Isak Roth and Dylan Manfredi for outstanding research assistance.

1 Introduction

Replication lies at the heart of the process by which science accumulates knowledge. The ability of other scientists to replicate an experiment or analysis demonstrates robustness, guards against false positives, puts an appropriate burden on scientists to make replication easy for others to do, and can expose the various "researcher degrees of freedom" like p-hacking or forking (Bavel et al., 2016; Begley and Ioannidis, 2015; De Vries et al., 2006; Gelman and Loken, 2013; Ioannidis et al., 2001; Ioannidis, 2005; Ioannidis et al., 2011, 2014; Koch and Jones, 2016; Lindsay, 2015; Martinson et al., 2005; Munafò et al., 2017; Nosek et al., 2015; O'Boyle et al., 2017; Open Science Collaboration, 2015; Silberzahn et al., 2018; Simmons et al., 2011; Simonsohn et al., 2014).

The most basic type of replication is "direct" replication, which strives to reproduce the creation or analysis of data using methods as close to those used in the original science as possible (Simons, 2014).

Direct replication is difficult and sometimes thankless. It requires the original scientists to be crystal clear about details of their scientific protocol, often demanding extra effort years later. Conducting a replication of other scientists' work takes time and money, and often has less professional reward than original discovery.

Because direct replication requires scarce scientific resources, it is useful to have methods to evaluate which original findings are likely to replicate robustly or not. Moreover, implicit subjective judgments about replicability are made during many types of science evaluations. Replicability beliefs can be influential when giving advice to granting agencies and foundations on what research deserves funding, when reviewing articles which have been submitted to peer-reviewed journals, during hiring and promotion of colleagues, and in a wide range of informal "post-publication review" processes, whether at large international conferences or small kaffeeklatches.

The process of examining and possibly replicating research is long and complicated. For example, the publication of Rand et al. (2012) resulted in a series of replications and subsequent replies (Bouwmeester et al., 2017; Rand, 2017; Rand et al., 2013; Tinghög et al., 2013). The original findings were scrutinized in a thorough and long process that yielded a better understanding of the results and their limitations. Many more published findings would benefit from such examination. The community is in dire need of tools that can make this work more efficient. Statcheck (Nuijten et al., 2016) is one such framework that can automatically identify statistical errors in finished papers. In the same vein, we present here a new tool to automatically evaluate the replicability of laboratory experiments in the social sciences.

There are many potential ways to assess whether results will replicate. We propose a simple, black-box, statistical approach, which is deliberately automated in order to require little subjective peer judgment and to minimize costs. This approach leverages the hard work of several recent multi-investigator teams who performed direct replications of experiments in psychology and economics (Camerer et al., 2016; Ebersole et al., 2016; Klein et al., 2014; Open Science Collaboration, 2015). Based on these actual replications, we fit statistical models to predict replication and analyze which objective features of studies are associated with replicability.

We have 131 direct replications in our dataset. Each can be judged categorically by whether it succeeded or failed, by a pre-announced binary statistical criterion. The degree of replication can also be judged on a continuous numerical scale, by the size of the effect estimated in the replication compared to the size of the effect in the original study. Our method uses machine learning to predict outcomes and identify the characteristics of study-replication pairs that can best explain the observed replication results (Camerer et al., 2018a; Hastie et al., 2009; Nave et al., 2018; Yarkoni and Westfall, 2017).

We divide the objective features of the original experiment into two classes. The first contains the statistical design properties and outcomes: among these features we have sample size, the effect size and p-value originally measured, and whether a finding is an effect of one variable or an interaction between multiple variables. The second class is the descriptive aspects of the original study which go beyond statistics: these features include how often a published paper has been cited and the number and past success of authors, but also how subjects were compensated. Furthermore, since our model is designed to predict the outcome of specific replication attempts we also include similar properties about the replication that were known beforehand.We also include variables that characterize the difference between the original and replication experiments — such as whether they were conducted in the same country or used the same pool of subjects. See Table S1 for a complete list of variables, and Tables S2–S11 for summary statistics.

The statistical and descriptive features are objective. In addition, for a sample of 55 of the study-replication pairs we also have measures of subjective beliefs of peer scientists about how likely a replication attempt was to result in a categorical Yes/No replication, on a 0-100% scale, based on survey responses and prediction market prices (Camerer et al., 2016; Dreber et al., 2015). Market participants predicted replication with an accuracy of 65.5%

Our proposed model should be seen as a proof-of-concept. It is fitted on an arguably too small data set with an indiscriminately selected feature set. Still, its performance is on par with the predictions of professionals, hinting at a promising future for the use of statistical tools in the evaluation of replicability.

2 Methods and Data

The data are combined from four replication projects, The Reproducibility Project in Psychology (RPP; Open Science Collaboration, 2015), the Experimental Economics Replication Project (EERP; Camerer et al., 2016) and Many Labs (ML) 1 and 3 (Ebersole et al., 2016; Klein et al., 2014). In most cases, one specific statistical test from each paper was selected for replication, but four papers had multiple effects replicated. In RPP and EERP, each experiment was replicated once. In the Many Labs projects all participating labs replicated every experiment and the final results were calculated from the pooled data. A total of 144 effects were studied.[1] After dropping observations with missing variables, our final dataset contains 131 study-replication pairs. For 55 of these observations we also have data from prediction markets prices (Camerer et al., 2016; Dreber et al., 2015).

[1] 100 RPP experiments, 16 from ML1, 10 from ML3, 18 from EERP.

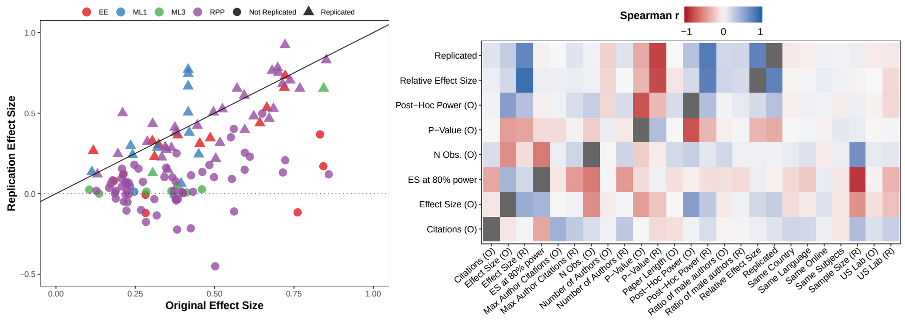

Figure 1: **Effect sizes and correlations**: (A) A plot of effect sizes (r) in each study pair. Data source is coded by color. Symbol shape denotes whether a study replicated (binary measure). Most points are below the 45-degree line, indicating that effect sizes are smaller in replications. Replications with a negative effect size have effects in the opposite direction compared to the original study. (B) A heatmap showing Spearman rank-order correlations between variables. Y-axis shows most important features with the two dependent variables on the top. `O` and `R` correspond to original and replication studies respectively. Most correlations are weak. See Table S1 for variable definitions and Figure S1 for a full correlation plot.

Dependent variables

There is no single best method for measuring whether a replication result is a failure or not. An active literature studies different strategies to evaluate replicability (see e.g. Andrews and Kasy, 2017; Leek et al., 2015; Simonsohn, 2015). For this paper, we have chosen to prioritize simplicity and focus on two measures, one binary and one continuous:

$$\text{Replicated} = \begin{cases} 1 & p_{\text{replication}} \leq 0.05 \text{ and effect in same direction} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Relative Effect Size} = \frac{\text{replication effect size } (r)}{\text{original effect size } (r)}$$

The binary model defines replication success as a statistically significant ($p \leq 0.05$) effect in the same direction as in the original study. This measure has often been criticized and is indeed overly simplistic. We use it since it can be compared to the prediction market estimates from previous studies (where subjects traded bets using the same measure).[2] The continuous model estimates the relative standardized effect size of the replication when compared to the original effect. It yields a more fine-grained notion of replication that does not depend on the peculiarities of hypothesis testing.[3]

[2] According to this definition, 56 replications are successful and 75 fail. 22 replication had effects going in the opposite direction compared to the original, but all with p-values larger than 0.05.

[3] By definition of a correlation coefficient, both the numerator and denominator are bounded by −1 and 1. In our data, the relative effect size varies between −0.9 and 2.38 with a mean of 0.49. As can be seen in the left plot of Figure 1, most relative effect sizes close to 0 are also “failed replications” in terms of the binary metric.

Independent variables

For each original-replication pair we have collected a large set of variables (see Figure 1B for the variable names or Table S1 for descriptions). The feature set includes objective characteristics of the original experiment, but also information about the replication that was known *before* it was carried out. We intentionally provide no theoretical justification for the inclusion of any specific feature, but simply gathered as many variables as possible. We leave it to the user of the model to decide which of these variables are relevant for their specific implementation, and provide information about the relative importance of each feature in Figure 4.

The heatmap of Spearman rank-order correlation coefficients in Figure 1B shows some correlation between our two outcomes and other features (the two top rows). Most relationships are weak. Ex ante expected correlations are strong (e.g., sample size and p-value) but not many other relationships are evident visually (see Figure S1 for a full correlation plot). Original effect sizes are correlated with binary replication and so are p-values, with Spearman ρ of 0.26 and 0.38 respectively.

Model Training

We use cross validation to avoid overfitting. To simultaneously evaluate variability of the accuracy metric[4] we nest two cross validation loops, as shown in Figure 2. In the inner loop, we search and validate algorithm-specific hyperparameters. Each such optimally configured model is then tested on 20% of the data in the outer loop. Our limited sample size forces us to use these validation sets for both reported performance statistics and algorithm selection (Figures 3 and S3). Because we make decisions based on these performance statistics, also our cross validated measures may suffer from some overfitting. We therefore evaluate pre registered predictions of the results of Camerer et al. (2018b) as a supplement.

When training the binary classification models, we do so with the goal to maximize the area under the curve (AUC) of a receiver operating characteristics (ROC) curve (Bradley, 1997). The metric accounts for the trade-off between successfully predicting positive and negative results respectively. Maximizing accuracy might result in a model that always predicts failure to replicate, and accurately predicts all unsuccessful replications, but incorrectly classifies all those that do replicate. The model with the highest AUC will instead be the one that minimizes the effects of this trade off, achieving high prediction rates for both positive and negative results simultaneously. The models predicting relative effect size are trained to minimize the mean squared prediction error.

We compare a number of popular machine learning algorithms (see Figure S3) and find that a Random Forest (RF) model has the highest performance. The outcome predicted by an RF algorithm is the result of averaging over a "forest" of decision trees. Each tree is fitted using a random subset of variables, and employs a hierarchical sequence of cutoffs to predict observations (Breiman, 2001). A simple tree with depth 2 might fit 0-1 replication success by first dividing cases by if sample size is below a cutoff, then, at each of those two

[4] Which is a potential issue since the small sample size could make predictions change depending on which observations are held out.

branches, by whether the original effect is a main effect or an interaction. Each end node is a prediction of the outcome variable. The algorithm is popular because it performs well without much human supervision.

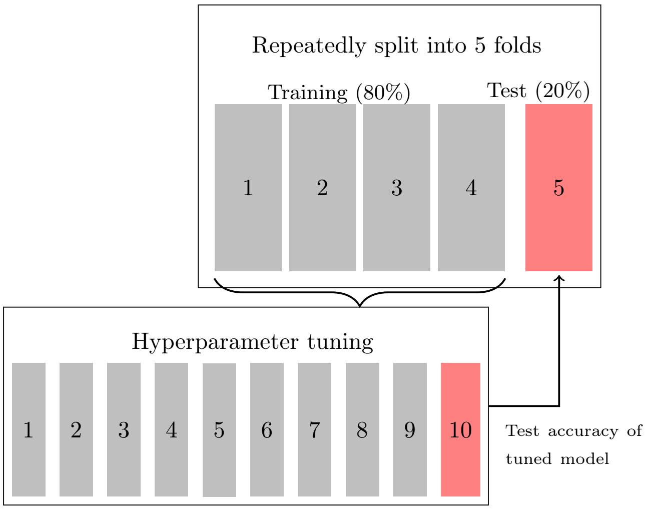

Figure 2: **Model Training, Nested Cross Validation (CV)**: First, the data is split into five parts. Four parts are used for training. For each model a 10-fold CV is run on this training data to find optimal hyperparameters for each algorithm. When training the LASSO, different values for λ (penalty to weakly correlated variables) are tested, for Random Forest the number of randomly selected features to consider at each split changes. In each run the model is trained on 9/10th of the data and tested on the last decile. The best version (with highest AUC) is trained on all of the training data and its accuracy is estimated on the fifth fold of the outer loop. The process is repeated with a different outer fold held out. After five runs, a new set is drawn, and the process is repeated until 100 accuracy metrics have been generated.

3 Results

The Random Forest model trained on the full feature set predicts binary replication with a median AUC of 0.79 (median accuracy of 69% at the 50% probability threshold), shown in the top bars of Figure 3. The bar width is the interquartile range of 100 performance resamples produced by the nested cross validation. We do not show confidence intervals since these resamples are not independent. The median is depicted as a dot.[5]

[5] Note that because of sampling error and other statistical properties, the upper bound for ideal replication forecasting is less than 100%; rather we would expect it to be around 90% (see Section S3). Why? Consider an artificial sample, measuring a, by construction, genuine effect with tests that have 90% power to detect it. A perfect model predicting significance in a second sample will only be right nine out of ten times. This theoretical ceiling is important since we should arguably normalize the distance between random guessing and the best possible level of prediction. For 69% accuracy, the normalized improvement over a random guess (50%) to perfection is $\frac{69-50}{100-50} = 0.38$. However using a more accurate upper bound of

The model predicting Relative Effect Size achieves a median R^2 of 0.19. The predicted and actual effect sizes have a median Spearman correlation of 0.36.

In the pre-registered validation of SSRP, 71% of binary replications are predicted correctly (AUC: 0.73). Relative effect size is predicted with a mean absolute prediction error of 0.43 and a Spearman correlation of 0.25. Individual predictions are presented in Figure 5.

A qualitative assessment of these results can be made in both relative and absolute terms.

First, classifier performance is substantially higher than that of a random model (which by definition has an AUC of 0.5), and is more accurate than a linear model using the same features (the last bar in each subplot of Figure 3, median AUC = 0.72). Furthermore, the continuous Random Forest (RF) model explains 19% of the variation in relative effect size, compared to the OLS R^2 of 0.06. The rank-order correlation coefficient of 0.36 (OLS: 0.27) between predicted and actual values is higher than the 0.21 correlation between original and relative effect size, indicating a fairly substantive performance improvement over a very simple heuristic.

When cross validated accuracy is compared to out of sample tests, the binary classifier achieves similar results. The continuous model performs worse, with a Spearman correlation coefficient that is 30% smaller.

Second, these machine-learned predictions, based on objective features, can be compared to subjective beliefs of replicability based on prediction market prices in earlier studies where social scientists traded on the probability of replication success. The market predicted 65.5% of the replications in our data correctly[6] (with an AUC of 0.73). While the model fares slightly better in this data set, it is difficult to draw conclusions on the relative performances of the methods based on such a small sample.

Predictive Power

In Figure 3, we further compare the performance of models in which certain classes of variables have been excluded. The observation of similar patterns for both sets of models is not surprising, given the high correlation of the two outcome measures (Spearman $\rho = 0.79$).

For both replication measures, the second bar shows that removing the dummy variables encoding the discipline of the study (Economics, Social Psychology or Cognitive Psychology) has little bearing on the results. The 64 Social Psychology replications have smaller effect sizes (mean of 0.33 compared to 0.47 for cognitive psychology), slightly larger p-values (0.017 compared to 0.01). Inbar (2016) argues that the association between contextual sensitivity (as measured on a scale from 0-5) and replicability found by Bavel et al. (2016) is spuriously identified from the difference in replication rates between fields. We show that many other variables also mediate these differences. For example,

90%, it is $\frac{69-50}{90-50} = 0.48$.

[6] Accuracy is calculated assuming that market prices reflect replication probabilities (Wolfers and Zitzewitz, 2006) and using a threshold of 0.5. We only have model predictions and market prices for 55 observations, including data from Dreber et al. (2015) with an accuracy rate of 68% and Camerer et al. (2016) where accuracy was 61%. In two follow-up papers, prediction markets perform better (Camerer et al., 2018b; Forsell et al., 2018). Pooling results across all four papers yields a prediction market accuracy rate of 73% (76/104).

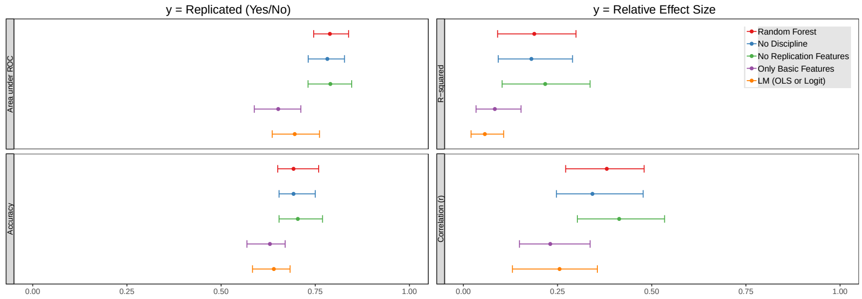

Figure 3: Interquartile range (IQR) and median of Random Forest classifier (left) and regression (right) out of sample performance. For the classifier, the optimal model (first from top) has an average AUC of 0.79 and accuracy of 70% at the 50% probability cutoff (accuracy is mainly driven by a high true negative rate; failed replications are predicted with an accuracy of 80%, while successful only with 56%). The optimal regression model has a median R^2 of 0.19 and a Spearman ρ of 0.38. The second bar from the top in each subplot shows unchanged model performance when dummy indicators for discipline (Economics, Social or Cognitive Psychology) are removed. The third has excluded any features unique to the replication effort (e.g. replication team seniority) with no observable loss of performance. The less accurate fourth model is only based on original effect size and p-value. Last, the model at the bottom is a linear model trained on the full feature set, for reference. See Figure S3 for more models.

by construction, holding sample size constant, interaction effects will have lower statistical power. Included social psychology experiments test interaction effects almost twice as often (44% vs 27% in cognitive psychology). If studies of interactions do not increase sample size appropriately, replicability will be lower.

The third bar shows no reduction in accuracy for a model in which all replication-specific features are excluded. The reason is likely that replication characteristics were standardized between experiments. No replication is conducted with a really small sample size, for example.

The fourth bar uses only original effect size and p-value. The decrease in accuracy shown in this bar implies that also other features are informative.

Feature Importance

The previous section summarized the *general* accuracy of the models, using different feature subsets. This section explores which objective features of experimental designs and results correlate with replicability, extending the analysis in RPP (Open Science Collaboration, 2015) with more variables and a larger data set.

The action-packed Figure 4 reports two metrics of feature importance for both the binary (blue) and continuous (red) models. The length of horizontal bars represents Random Forest variable importance, a measure of the relative frequency of each feature in the many individual decision trees. Features that are included in a large proportion of the individual trees will have a long bar. For example, the top three bars are the power, p-value, and effect size in the original studies. The variables are sorted by their importance in predicting binary classification.

Since the RF model is hierarchical and nonlinear, a single variable can be included in many different individual trees with both positive and negative effects on predicted outcome. While we can identify the most important variables in the model, we cannot determine the direction of their influence. We therefor also present the *linear* effect of each variable in a Logit model. These are shown in small boxes between the variable names (on the left) and the bars on the right. This analysis uses only variables that have been selected as important by a LASSO.[7]

Pre-registered Out of Sample Validation

The Social Sciences Replication Project (SSRP; Camerer et al., 2018b) replicated 21 systematically selected papers published in *Nature* and *Science* published between 2011 and 2015. The authors also collected beliefs through a survey and a prediction market. We registered the predictions of the model before the replications had been conducted.[8] The results from this out of sample test are summarized and compared to market and survey beliefs in Figure 5

The out of sample predictions achieve accuracy similar to the median cross validated level, at 71% (AUC: 0.73).[9] When compared to researcher beliefs, the model has a mean absolute prediction error of 0.43, while the market achieves 0.30 and the survey 0.35. The difference between model and market is significant (Wilcoxon signed-rank test, $z = -2.52$, $p = 0.012$, $n = 21$), however more data is needed to verify these differences.

The model predicts relative effect size with a Spearman correlation of 0.25 ($p = 0.274$), lower than the cross validated measure of 0.38. The mean absolute deviation is 0.33. A Wilcoxon sign-rank test cannot reject that the distributions of predicted and actual effect sizes are the same, $z = -1.00$, $p = 0.317$.

Results are summarized in Figure 5. We see that the model produces quite conservative forecasts of effect size, often closer to 0.5 than the actual outcome. This results in large errors whenever the actual effect is substantially different from the original, which is most often true for failed replications. While the market and survey perform better than the model in this sample, the plot shows

[7] The LASSO is a regularization algorithm, minimizing squared errors (or deviance) while keeping the absolute value of coefficients constrained by a penalty term. This method tends to shrink estimated coefficients that are unimportant towards zero, removing some variables completely (Hastie et al., 2009). For the many variables that are common in the RF trees but have zero LASSO weights, there are blank spaces between variable definitions and RF-frequency bars. The features selected with positive weight by LASSO are then re-fitted in a regular Logit model (to "unshrink" their weights) and the coefficients of that non-regularized model are presented Figure 4.

[8] Available at: `https://osf.io/w2y96`

[9] The replications were conducted in a two-stage procedure, where more data was collected if the results from the first phase were not significant. Here, we use the results from the pooled data. If the model predicted a successful Stage 1 replication these predictions are used. If it predicted a failed first stage, predicted effect size and replication probability from Stage 2 are used instead. In the Supplementary Material we provide the results when only Stage 1 is used. The results are similar.

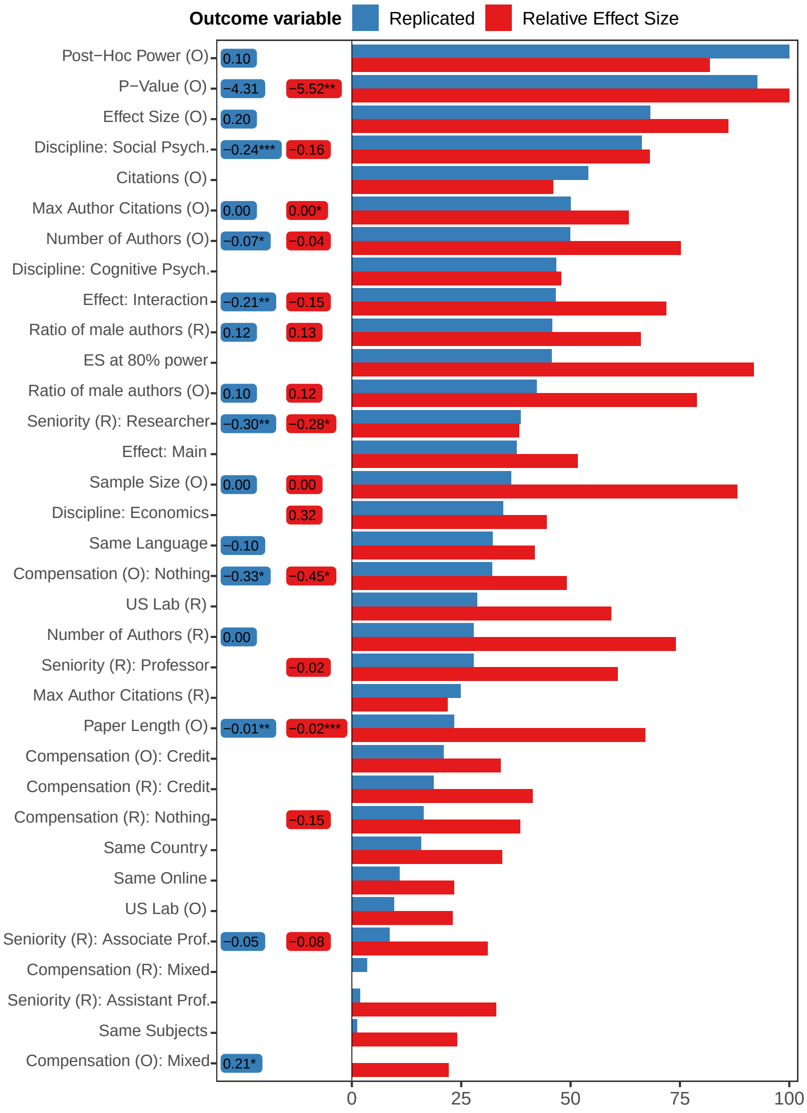

Figure 4: Right side contains variable importance measures for all features used in the Random Forest, for both regression (red) and classification (blue) models, sorted by decreasing contribution to the predictive power of the classifier. To the left are average marginal effects for those variables selected by a LASSO and then re-fit in a Logistic model. Predictably, most of the top variables are statistical properties related to replicability and publication, but also other variables seem to be informative, especially for the Random Forest. For example, whether or not the effect tested is an interaction effect, as well as the number of citations are important. Last, note that the two top variables are basically non-linear transformations of one another. Stars indicate significance: $p \leq 0.01(***)$, $p \leq 0.05(**)$, and $p \leq 0.1(*)$.

how the measures often yield the same prediction. When they do not, it is often because the model incorrectly predicts that a replication will fail.

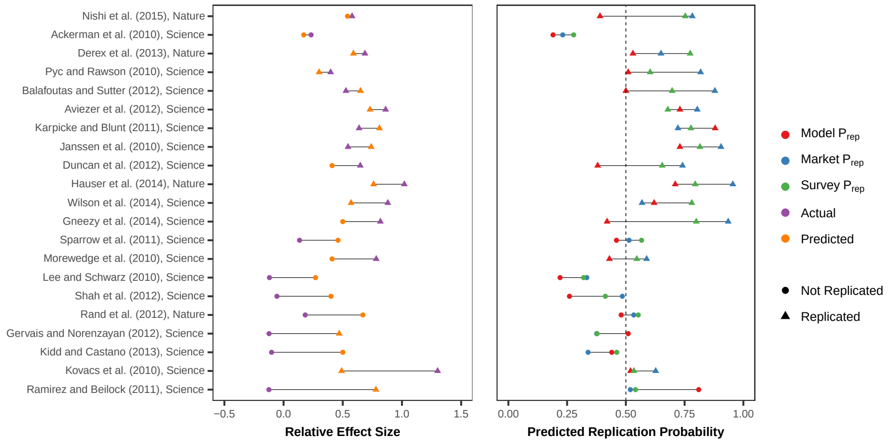

Figure 5: Predicted and actual results of the SSRP. Model predictions were registered before the experiments had been conducted. The left panel shows predicted relative effect size in orange and actual in purple, sorted by increasing prediction error. Right panel shows replication probability as predicted by the model, a prediction market, and a survey respectively. Data points are represented by a triangle when the replication was successful ($p < 0.05$ and an effect in the same direction). In the left panel, the shape of purple data points are based on the predicted replication probability instead, using a 50% probability cutoff.

4 Discussion

We have derived an automated, data-driven method for predicting replicability of experiments. The method uses machine learning to discover which features of studies predict the strength of actual replications. Even with our fairly small data set, the model can forecast replication results with substantial accuracy — around 70%.

Predictive accuracy is sensitive to the variables that are used, in interesting ways. The statistical features (p-value and effect size) of the original experiment are the most predictive. However, the accuracy of the model is also increased by variables such as the nature of the finding (an interaction, compared to a main effect), number of authors, paper length and the lack of performance incentives. All those variables are associated with a reduction in the predicted chance of replicability.

The third bar in Figure 3 showed unchanged performance for a model with all replication-specific features excluded. There are a couple of possible reasons why removing replication features has no impact on model performance. For one thing, most variables have a small impact, that would be easier to identify in a larger data set. Second, it is obviously the case that a larger planned sample size has a direct impact on replication probability[10]. The reason why we do not find such a relationship is probably because our data mainly consists of well-powered replications and do not include multiple replications of the same experiment with different sample sizes. This makes it hard for the model to capture any variation in replicability caused by changes in planned sample size. It is also possible that the model is unable to separate the increase in power from the fact that weaker effects required larger replication samples.

The fourth bar in Figure 3 presents the accuracy of a simple model that is only trained on effect size and p-value of the original experiment. It is not quite as accurate as models with more features, but still on par with the linear model trained on the full feature set. The analysis of correlations in Open Science Collaboration (2015) indicated the opposite, that experience of the experimenters and other such features are unimportant. With the substantial variability in out-of-sample accuracy, it is difficult to say for sure, but our results do indicate that these other features are correlated and indeed contribute to higher accuracy.

We now probe a bit further into three results.

The first result is that one variable that is predictive of poor replicability is whether central tests describe interactions between variables or (single-variable) main effects. Only eight of 41 interaction effect studies replicated, while 48 of the 90 other studies did. As Figure 4 shows, the interaction/main effect variable is in the top 10 in RF importance and is predictive, for both the binary and continuous replication measures.

There is plenty of room for reasoned debate about the validity of apparent interactions. Here is our view: Interactions are often slippery statistically because detecting them is undermined by measurement error in either of two variables. In early discussion of p-hacking it was also noted that studies which hoped to find a main effect often end up concluding that there is a main effect which is only significant in part of the data (i.e., an interaction effect). The lower replication rates for interaction effects might be spurious, however. Tests of interactions often require larger samples, which could mean that the replications of these studies have lower power relative to those studies evaluating non-interacted effects. Nonetheless, the replicability difference is striking and merits further study. It is possible that the higher standard of evidence for establishing interactions needs to be upheld more closely.

The second result is that some features that vary across studies are *not* robustly associated with poor replication: These include measures of language, location and subject type differences between replication and original experiments, as well as most of the variation in compensation (except for having no compensation at all, which is correlated with lower replicability).

Our third result is that the model performs on par with previously collected peer judgments (subjective beliefs as measured by prediction market prices). In the sample used to estimate the model, it performed somewhat better than the prediction market, although we only had prediction market data on a subset of $n = 55$ studies. On the other hand the prediction market performed better than the model on the out of sample prediction test, but this was based

[10] With higher power, it follows that there is a higher probability of rejecting a false null hypothesis, and thus also the corresponding probability of replicating a true result. See e.g. Gelman and Carlin (2014) for a discussion of replication power.

on a small sample of $n = 21$. More data is needed to compare statistical approaches with peer judgments in prediction markets and surveys, to test which approach is associated with the most accurate predictions, and to look for potential complementarities. If the goal is replication prediction, the model has logistical advantages compared to running prediction markets, which require both participants and costly monetary incentives.

Studying the differences between our algorithmic predictions and expert scientific judgment adds to a long literature comparing machine and man. For at least seventy years, it has been known that in many domains of professional judgment, simple statistical models can predict complex outcomes — PhD success, psychiatric disorders, recidivism, personality — as accurately as humans do subjectively (Bishop and Trout, 2004; Dawes, 1979; Meehl, 1954; Youyou et al., 2015). Today, with the tremendous increase in data availability and development of more sophisticated predictive models, statistical prediction has become useful in many new areas (e.g., Kleinberg et al., 2018). It is likely that in some form, statistical methods will also increase the quality of human evaluation and prediction of scientific findings. The results presented in this paper suggests that there could be room for statistical methods to aid researchers when reviewing their peers' experiments.

Applications

Our method could be used in pre- and post-publication assessment, preferably after a lot more replication evidence is available to train the algorithm. In the current mainstream pre-publication review process, the decision about whether to publish a paper is almost entirely guided by the opinions of a small set of peer scientists and an editor. A systematic, fast, and accurate numerical method to estimate replicability could add more information in a transparent and fair way to this process. For example, when a paper is submitted an editorial assistant can code the features of the paper, plug those features into the models, and derive a predicted replication probability. This number could be used as one of many inputs helping editors and reviewers to decide whether a replication should be conducted before the paper is published.

Post-publication, the model could be used as an input to decide which previously published experiments should be replicated. The criteria should depend on the goal of replication efforts. If the goal is to quickly locate papers unlikely to replicate, then papers with low predicted replicability should be chosen.

Since replication is costly and laborious, using predicted probability can guide scarce resources toward where they are most scientifically useful.

An important concern with any predictive algorithm is that its application will likely change incentives, and impact how scientists design their studies, undermining the algorithm's value. Some of these "corrupting" (Campbell, 1979) effects will actually be good: For example, since testing interaction effects seem to negatively associated with predicted replicability, scientists may be motivated to avoid searching for such interactions. But that could be an improvement, if such effects are difficult to find robustly with sample sizes used previously. Alternatively, scientists who are keen to find interactions can use higher-powered designs, which will increase predicted replicability.

Other changes in practice to “game” the algorithm will likely be harmless, and some changes could reduce predictive accuracy. For example, longer papers tend to replicate less well. If scientists all shorten their papers (to increase their predicted replicability), without changing the quality of the science, then the paper length variable will gradually lose diagnostic value. Any implementation will need to anticipate this type of gaming.

Some types of “gaming” could be truly unwanted. The trade off between algorithmic fairness and accuracy is a highly important question that is currently being studied extensively. In our case, including the gender composition or seniority of the author team potentially increases the risk that the model is discriminatory. If needed, such variables could easily be removed, with only a small penalty to accuracy. However, excluding a variable like gender composition will not necessarily remove the model’s tendency to discriminate, as this variation could be captured also in other features (Kleinberg et al., 2016).

Of course, there are limits to how much we can conclude from our results. The data we use is not representative for experimental social science — the accuracy level and variable importance statistics may be specific to our sample, or to psychology and economics. Our sample of studies is also very small; having more actual replications is crucial to ensure that the model functions robustly.

Moreover, the correlations we find do not identify causal mechanisms, so changing research practices (as in the “gaming” scenarios above) may have unknown consequences. Rather, our model is theory agnostic by design. We aim to predict replicability, not understand its causes. The promising and growing literature taking a theoretical approach to this questions (see e.g. Andrews and Kasy, 2017; Simonsohn, 2015; Simonsohn et al., 2014) should be seen as a complement to our work and could hopefully be used to improve future versions of this predictive model. Simultaneously, our insights will hopefully be useful for future theoretical investigations.

The future is bright. There will be rapid accumulation of more replication data, more outlets for publishing replications (Simons et al., 2014), new statistical techniques, and — most importantly — enthusiasm for improving replicability among funding agencies, scientists, and journals. An exciting replicability “upgrade” in science, while perhaps overdue, is taking place.

References

Andrews, I. and M. Kasy (2017). “Identification of and Correction for Publication Bias.” *NBER Working Paper Series (No. 23298).*

Bavel, J. J. V. et al. (2016). “Contextual Sensitivity in Scientific Reproducibility.” *Proceedings of the National Academy of Sciences* 113.23, pp. 6454–6459.

Begley, C. and J. P. Ioannidis (2015). “Reproducibility in Science.” *Circulation Research* 116.1, pp. 116–126.

Bishop, M. A. and J. D. Trout (2004). *Epistemology and the Psychology of Human Judgment.* Oxford University Press.

Bouwmeester, S. et al. (2017). “Registered Replication Report: Rand, Greene, and Nowak (2012).” *Perspectives on Psychological Science* 12.3, pp. 527–542.

Bradley, A. P. (1997). “The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms.” *Pattern Recognition* 30.7, pp. 1145–1159.

Breiman, L. (2001). “Random Forests.” *Machine Learning* 45.1, pp. 5–32.

Camerer, C. F., G. Nave, and A. Smith (2018a). “Dynamic Unstructured Bargaining with Private Information: Theory, Experiment, and Outcome Prediction via Machine Learning.” *Management Science.*

Camerer, C. F. et al. (2016). “Evaluating Replicability of Laboratory Experiments in Economics.” *Science* 351.6280, pp. 1433–1436.

Camerer, C. F. et al. (2018b). “Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015.” *Nature Human Behaviour* 2.9, pp. 637–644.

Campbell, D. T. (1979). “Assessing the impact of planned social change.” *Evaluation and Program Planning* 2.1, pp. 67–90.

Dawes, R. M. (1979). “The Robust Beauty of Improper Linear Models in Decision Making.” *American Psychologist* 34.7, pp. 571–582.

De Vries, R., M. S. Anderson, and B. C. Martinson (2006). “Normal Misbehavior: Scientists Talk about the Ethics of Research.” *Journal of Empirical Research on Human Research Ethics* 1.1, pp. 43–50.

Dreber, A. et al. (2015). “Using Prediction Markets to Estimate the Reproducibility of Scientific Research.” *Proceedings of the National Academy of Sciences* 112.50, pp. 15343–15347.

Ebersole, C. R. et al. (2016). “Many Labs 3: Evaluating Participant Pool Quality across the Academic Semester via Replication.” *Journal of Experimental Social Psychology* 67, pp. 68–82.

Forsell, E. et al. (2018). “Predicting Replication Outcomes in the Many Labs 2 Study.” *Journal of Economic Psychology.*

Gelman, A. and J. Carlin (2014). “Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors.” *Perspectives on Psychological Science* 9.6, pp. 641–651.

Gelman, A. and E. Loken (2013). “The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No “fishing Expedition” or “p-Hacking” and the Research Hypothesis Was Posited Ahead of Time.”

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Second. Springer Series in Statistics. Springer.

Inbar, Y. (2016). “Association between Contextual Dependence and Replicability in Psychology May Be Spurious.” *Proceedings of the National Academy of Sciences* 113.34, E4933–E4934.

Ioannidis, J. P. et al. (2001). “Replication Validity of Genetic Association Studies.” *Nature Genetics* 29.3, pp. 306–309.

Ioannidis, J. P. A. (2005). “Why Most Published Research Findings Are False.” *PLOS Medicine* 2.8, e124.

Ioannidis, J. P. A., R. Tarone, and J. K. McLaughlin (2011). “The False-Positive to False-Negative Ratio in Epidemiologic Studies.” *Epidemiology* 22.4, pp. 450–456.

Ioannidis, J. P. A. et al. (2014). “Publication and Other Reporting Biases in Cognitive Sciences: Detection, Prevalence, and Prevention.” *Trends in Cognitive Sciences* 18.5, pp. 235–241.

Klein, R. A. et al. (2014). “Investigating Variation in Replicability: A “Many Labs” Replication Project.” *Social Psychology* 45.3, pp. 142–152.

Kleinberg, J., S. Mullainathan, and M. Raghavan (2016). “Inherent Trade-Offs in the Fair Determination of Risk Scores.”

Kleinberg, J. et al. (2018). “Human Decisions and Machine Predictions.” *The Quarterly Journal of Economics* 113.1, pp. 237–293.

Koch, C. and A. Jones (2016). “Big Science, Team Science, and Open Science for Neuroscience.” *Neuron* 92.3, pp. 612–616.

Leek, J. T., P. Patil, and R. D. Peng (2015). “A Glass Half Full Interpretation of the Replicability of Psychological Science.” *arXiv:1509.08968 [stat]*.

Lindsay, D. S. (2015). “Replication in Psychological Science.” *Psychological Science* 26.12, pp. 1827–1832.

Martinson, B. C., M. S. Anderson, and R. de Vries (2005). “Scientists Behaving Badly.” *Nature* 435, pp. 737–738.

Meehl, P. E. (1954). *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* Minneapolis, MN, US: University of Minnesota Press.

Munafò, M. R. et al. (2017). “A Manifesto for Reproducible Science.” *Nature Human Behaviour* 1.1, p. 0021.

Nave, G. et al. (2018). “Musical Preferences Predict Personality: Evidence From Active Listening and Facebook Likes.” *Psychological Science* 29.7, pp. 1145–1158.

Nosek, B. A. et al. (2015). “Promoting an Open Research Culture.” *Science* 348.6242, pp. 1422–1425.

Nuijten, M. B. et al. (2016). “The Prevalence of Statistical Reporting Errors in Psychology (1985–2013).” *Behavior Research Methods* 48.4, pp. 1205–1226.

O’Boyle, E. H., G. C. Banks, and E. Gonzalez-Mulé (2017). “The Chrysalis Effect: How Ugly Initial Results Metamorphosize Into Beautiful Articles.” *Journal of Management* 43.2, pp. 376–399.

Open Science Collaboration (2015). “Estimating the Reproducibility of Psychological Science.” *Science* 349.6251.

Rand, D. G. (2017). “Reflections on the Time-Pressure Cooperation Registered Replication Report.” *Perspectives on Psychological Science* 12.3, pp. 543–547.

Rand, D. G., J. D. Greene, and M. A. Nowak (2012). “Spontaneous Giving and Calculated Greed.” *Nature* 489.7416, pp. 427–430.

— (2013). “Rand et Al. Reply.” *Nature* 498.7452, E2–E3.

Silberzahn, R. et al. (2018). “Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results.” *Advances in Methods and Practices in Psychological Science* 1.3, pp. 337–356.

Simmons, J. P., L. D. Nelson, and U. Simonsohn (2011). “False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science* 22.11, pp. 1359–1366.

Simons, D. J. (2014). “The Value of Direct Replication.” *Perspectives on Psychological Science* 9.1, pp. 76–80.

Simons, D. J., A. O. Holcombe, and B. A. Spellman (2014). “An Introduction to Registered Replication Reports at Perspectives on Psychological Science.” *Perspectives on Psychological Science* 9.5, pp. 552–555.

Simonsohn, U. (2015). “Small Telescopes Detectability and the Evaluation of Replication Results.” *Psychological Science* 26.5, pp. 559–569.

Simonsohn, U., L. D. Nelson, and J. P. Simmons (2014). “P-Curve: A Key to the File-Drawer.” *Journal of Experimental Psychology: General* 143.2, pp. 534–547.

Tinghög, G. et al. (2013). “Intuition and Cooperation Reconsidered.” *Nature* 498.7452, E1–E2.

Wolfers, J. and E. Zitzewitz (2006). “Interpreting Prediction Market Prices as Probabilities.” *NBER Working Paper Series (No. 12200).*

Yarkoni, T. and J. Westfall (2017). “Choosing Prediction over Explanation in Psychology: Lessons from Machine Learning.” *Perspectives on Psychological Science* 12.6, pp. 1100–1122.

Youyou, W., M. Kosinski, and D. Stillwell (2015). “Computer-Based Personality Judgments Are More Accurate than Those Made by Humans.” *Proceedings of the National Academy of Sciences* 112.4, pp. 1036–1040.