

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Estimating credibility of science claims
– Analysis of forecasting data from
metascience projects

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

At Massey University, Albany

New Zealand

Michael Gordon

2021

Abstract

The veracity of scientific claims is not always certain. In fact, sufficient claims have been proven incorrect that many scientists believe that science itself is facing a “replication crisis”. Large scale replication projects provided empirical evidence that only around 50% of published social and behavioral science findings are replicable. Multiple forecasting studies showed that the outcomes of replication projects could be predicted by crowdsourced human evaluators. The research presented in this thesis builds on previous forecasting studies, deriving new findings and exploring new scope and scale. The research is centered around the DARPA SCORE (Systematizing Confidence in Open Research and Evidence) programme, a project aimed at developing measures of credibility for social and behavioral science claims. As part of my contribution to SCORE, myself, along with a international collaboration, elicited forecasts from human experts via surveys and prediction markets to predict the replicability of 3000 claims. I also present research on other forecasting studies.

In chapter 2, I pool data from previous studies to analyse the performance of prediction markets and surveys with higher statistical power. I confirm that prediction markets are better at forecasting replication outcomes than surveys. This study also demonstrates the relationship between p-values of original findings and replication outcomes. These findings are used to inform the experimental and statistical design to forecast the replicability of 3000 claims as part of the SCORE programme. A full description of the design including planned statistical analyses is included in chapter 3. Due to COVID-19 restrictions, our generated forecasts could not be validated through direct replication, experiments conducted by other teams within the SCORE

collaboration, thereby preventing results being presented in this thesis. The completion of these replications is now scheduled for 2022, and the pre-analysis plan presented in Chapter 3 will provide the basis for the analysis of the resulting data.

In chapter 4, an analysis of ‘meta’ forecasts, or forecasts regarding field wide replication rates and year specific replication rates, is presented. We presented and published community expectations that replication rates will differ by field and will increase over time. These forecasts serve as valuable insights into the academic community’s views of the replication crisis, including those research fields for which no large-scale replication studies have been undertaken yet. Once the full results from SCORE are available, there will be additional insights from validations of the community expectations.

I also analyse forecaster’s ability to predict replications and effect sizes in Chapters 5 (Creative Destruction in Science) and 6 (A creative destruction approach to replication: Implicit work and sex morality across cultures). In these projects a ‘creative destruction’ approach to replication was used, where a claim is compared not only to the null hypothesis but to alternative contradictory claims. I conclude forecasters can predict the size and direction of effects.

Chapter 7 examines the use of forecasting for scientific outcomes beyond replication. In the COVID-19 preprint forecasting project I find that forecasters can predict if a preprint will be published within one year, including the quality of the publishing journal. Forecasters can also predict the number of citations preprints will receive.

This thesis demonstrates that information about scientific claims with respect to replicability is dispersed within scientific community. I have helped to develop methodologies and tools to efficiently elicit and aggregate forecasts. Forecasts about scientific outcomes can be used as guides to credibility, to gauge community expectations and to efficiently allocate sparse replication resources.

Acknowledgements

PhDs are always a journey. I am a different person, and the world is different place from when my journey started in 2019. This journey would not have been possible without many people who I would like to thank.

Firstly, thank you to my supervisor Professor Thomas Pfeiffer. I have really appreciated your guidance, feedback, and support over the last three years. You have been patient and encouraging, making me feel like a valued collaborator in the projects we did together. You have been a great mentor, teacher, and friend. I will miss working together. Thank you also to Dr. Adam Smith for your support and guidance from the beginning of my masters through to now. Panyse and Stan, I really enjoyed being a team with you two. Thank you for all the conversations, lunches, support and friendship. I am glad I didn't have to do this as the only PhD student.

To my parents, Kerry and Paul, thank you for your support, I don't know that I would have done this without your encouragement. You are both always so uplifting and calming. Thank you. Thank you also to the rest of my family – especially for acting interested. To my Granny who didn't get to see me complete this PhD, thank you for always being a proud and caring grandmother.

Leo, you came in the last year of my study and made it the best year.

Lastly and most importantly, thank you to my wife, Karissa. My constant supporter, the best listener and my biggest encourager. This would have not have been possible without you and the sacrifices you have made. I am eternally grateful.

Ephesians 3:20

Table of Contents

Abstract	iii
Acknowledgements	vi
1 Chapter 1 - Introduction	1
1.1 Replication Crisis.....	1
1.2 Using Human Forecasts to Evaluate Replicability.....	4
1.3 SCORE: Systematizing Confidence in Open Research Evidence.....	5
1.4 Thesis Structure	8
1.5 References.....	11
2 Chapter 2: Predicting replicability - analysis of survey and prediction market data from large-scale forecasting projects	16
2.1 Abstract.....	17
2.2 Introduction.....	18
2.3 Methods	19
2.4 Results.....	25
2.5 Discussion.....	35
2.6 References.....	37
3 Chapter 3: SCORE Pre-registration.....	41
3.1 Context.....	42
3.2 Methodology.....	43
3.3 Pre-registration for the statistical analysis of the meta-markets and meta-surveys...	47
3.4 Pre-registration for the statistical analysis of the regular markets and surveys	50
3.5 Scores.....	59
3.6 Amendments	65
3.7 Appendices	70
3.8 COVID19 – Claims pre-registration documentation for SCORE TA2/Team KeyW80	
3.9 References.....	92
4 Chapter 4: Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE program.....	93

4.1	Abstract.....	95
4.2	Introduction.....	95
4.3	Methods	97
4.4	Results.....	100
4.5	Discussion.....	104
4.6	References.....	106
4.7	Supplementary Material.....	108
5	Chapter 5: Creative destruction in science	112
5.1	Abstract.....	113
5.2	Forecasting Creative Destruction Replication Results.....	113
5.3	Supplements for “Creative Destruction in Science”	115
5.4	Supplement 9: Detailed Report of the Forecasting Results.....	122
5.5	References.....	133
6	Chapter 6: A creative destruction approach to replication: Implicit work and sex morality across cultures	134
6.1	Abstract.....	135
6.2	Forecasting Survey	135
6.3	Supplementary Materials	138
6.4	References.....	156
7	Chapter 7: Forecasting the future of Covid-19 preprints	157
7.1	Abstract.....	158
7.2	Introduction.....	159
7.3	Methods	161
7.4	Results.....	165
7.5	Discussion.....	172
7.6	References.....	174
8	Chapter 8: Conclusion.....	177
8.1	Findings, implications, and contributions to knowledge.....	177
8.2	Future area of study	184

8.3	Final thoughts	186
8.4	References.....	187
9	Appendices	192
9.1	Appendix 1: Systematizing Confidence in Open Research and Evidence (SCORE) 193	
9.2	Appendix 2: DRC16 Forms	34

List of Figures

Figure 2-1. Market Beliefs.....	28
Figure 2-2. Market Price and Replication Outcome Correlation and Meta-Analysis..	29
Figure 2-3. Market and Survey Correlations.....	30
Figure 2-4. Market Dynamics.....	34
Figure 2-5. P-Value and Replication Outcomes.....	35
Figure 4-1. A. Expected replication rate for publications from different 2-year periods. B. Expected replication rate for publications from different fields.....	102
Figure 4-2. A. In-field vs. out-of-field responses. Participants predict a higher replication rate for their fields of interest, as compared to other fields. B. Difference of evaluation of a field by in-field and out-field participants (in percent points).....	103
Figure 5-1 : Correlation between realized effect sizes and mean predicted effect sizes.....	124
Figure 6-1. Actual effect size vs average forecast (Cohen's d).....	148
Figure 7-1 – Forecasts of publication outcomes.....	167
Figure 7-2 – Forecasts of journals of publication.....	168
Figure 7-3 – Actual citation ranks vs. forecasted citation ranks.....	169
Figure 9-1. Relationships between research teams comprising the three technical areas (TAs) of the SCORE program.....	2
Figure 9-2. Model of a bushel claim set for a single paper.....	7
Figure 9-3. Overview of the IDEA protocol, as adopted in the repliCATS project.....	9
Figure 9-4. Overview of the Replication Markets workflow.....	12
Figure 9-5: Labeling spans for sample size, sample details, and subject compensation.....	16
Figure 9-6: Labeling spans for sample elements excluded and the reason they were excluded...	16
Figure 9-7: Labeling the sample, experimental methods employed, and factors under study	16

List of Tables

Table 2-1: Main features of individual projects.....	26
Table 3-1: Confidence Score Descriptors	59
Table 3-2: COVID19 Confidence Score Descriptions	88
Table 4-1. Descriptive statistics of survey and market forecasts.	101
Table 4-2. List of Journals and Discipline Clusters	108
Table 4-3. List of questions in the initial survey and market.....	110
Table 4-4. p-values for pairwise t-tests for time-specific responses (df = 225 for all tests)	110
Table 4-5. p-values for pairwise t-tests for topic-specific responses (df = 225 for all tests).....	111
Table 4-6. Relation between forecast for the overall replication rate in score and demographic characteristics.....	111
Table 5-1: Correlation between forecasted and observed effect sizes.	124
Table 5-2: Forecasts of interaction effects and moderators in terms of squared prediction error (Brier score).	126
Table5-3: Summary statistics of measures in the exploratory hypotheses.....	127
Table 5-4: Forecaster beliefs and demographics on squared prediction error (Brier Score).....	127
Table 5-5: Robustness test for hypothesis 1 for predictions on simple effects (1), interaction effects (2), and moderator effects (3) separately.	129
Table 5-6: Forecaster beliefs and demographics on squared prediction error (Brier Score) for predictions on simple effects, interaction effects and moderator effects separately.....	130
Table 5-7: Forecaster beliefs and demographics on squared prediction error (Brier Score) for main effect of candidate gender on male evaluators only.	131
Table 6-1. Association between forecasted and observed effect sizes.	147
Table 6-2: Forecasts of moderator effects relative to simple effects in terms of squared prediction error (Brier score).....	149
Table 6-3: Summary of the differences between meta-analyzed effect sizes and forecasts (standard errors in parenthesis)	149
Table 6-4: Regression estimates of accuracy on country indicators.	151
Table 6-5: P-values resulting from pairwise Wald tests on country coefficients shown in Table S6-4 being different from each other.	152
Table 6-6: Regression estimating the effects of academic seniority on forecasting accuracy. ..	153
Table 6-7: Forecasted and realized effect sizes separately for each major sample of participants.	155

Table 7-1 – Pairwise Pearson correlation coefficients of survey questions and outcomes	172
Table 9-1. Journals comprising the Common Task Framework (CTF)	202
Table 9-2: A glossary of key terms as they are used for the SCORE program	5
Table 9-3: A single claim trace of a paper is composed of four levels.	6
Table 9-4: Discourse classes used in semantic parsing for the A+ method (TwoSix)	14
Table 9-5. Forms of empirical credibility assessment.....	20

1 Chapter 1 - Introduction

This thesis consists of research regarding estimating the credibility of scientific claims by way of human forecasts for the outcomes of replication attempts. My research relates to the ‘replication crisis’ and to studying the processes of science, also known as metascience¹. Before providing a more detailed overview of each paper and their relations, I briefly introduce the replication crisis, human forecasting, and their intersection in the SCORE project.

1.1 Replication Crisis

In the past decade, concerns over the credibility of scientific research have been raised (1,2). The rate of reproducibility of claims made in scientific publications has been discussed in medicine (3–6), biology (7–10), computer science (11–13), marketing (14) and sports science (15). Particular scrutiny was directed at the social and behavioural sciences, including psychology (16–21), economics (22), philosophy (23) and social science (24), where large scale systematic replication projects provided empirical evidence that many findings do not ‘hold up’ under direct replication. The results of replication projects combined with theoretical concerns over false positives, questionable research practices (QRPs) and publication biases (2,25–28) led to a perception that there is a ‘replication crisis’ (29).

The concept of replication (see Table 9-2 and Table 9-5 for a glossary of key terms)— that is, verification of a scientific claim using independent evidence - is essential in science (30). However,

¹ Also known as science of science or philosophy of science

what constitutes a replication differs between academic fields and specific studies, as even the most similar experiments will inevitably have small differences (31). Nosek and Errington (32) define a replication as a theoretical commitment, where, if results are consistent with an original finding, then confidence in this finding is increased while conversely, if results are inconsistent with an original finding, confidence in this finding is decreased. This definition is wide and includes many different types of assessments of scientific claims, all of which can be measures of credibility² (Table 9-5). Most commonly, however, replication (or reproductions) are split into three categories; computation replication, direct replication and conceptual replication (29).

Computational replication is a repeat of the same analysis on the same data, a form of replication that should have a success rate of near 100%; however, one study showed only 70% of psychology papers were computationally reproducible (33). Computational replication can also be used to find ambiguities in the description of analytic procedures. Direct replication attempts to repeat the original methodology of a study, including the data collection process, on new subjects (30). Conceptual replications include intentional differences in sample or methodology to test the theoretical boundaries of a finding (29).

Direct replications are the most common form of replication used to assess the credibility of research across a field in large scale replication projects (21–24), with the exception of the Many Labs projects. The Many Labs projects sit between direct and conceptual replications, as they aim to test the variation in replicability across different samples and settings, using direct replication (16–20). There is argument for superiority of conceptual replications for scientific progress as it is the theory that is tested rather than a specific methodology (34). In addition to the standard

² Credibility here refers to the collection of assessments which provide supporting information (or refuting information) and includes tests of replicability, robustness, generalizability and data analytic reproducibility

definitions of replication, there is the ‘creative destruction’ approach, where hypotheses in similar spaces are simultaneously tested (35,36) in an attempt to use replications to replace or revise theories. Most of the research in this thesis, including the SCORE project relies on direct replications. Replication outcomes are most commonly discussed as a dichotomy – the replication was either successful or not. Success is usually determined by whether the replication identified a statistically significant finding in the same direction (37). However, other outcomes (such as effect sizes) are often reported in replication studies (19,21,24). There are benefits and disadvantages of using binary success criteria. Binarizing the replication outcome results in information loss such as the differences in effect sizes and the degree to which confidence intervals of the original and replication effects overlap. By extension, this method of evaluating replication speaks only to whether a claim replicates, rather than the degree to which it replicates. Conversely, the binary approach does provide some advantages. Binary outcomes provide a simple and easily definable and measurable definition of replication success. In addition, binary outcomes are often easier to forecast, and prediction markets (as used in this thesis) rely upon betting on a future event with a binary outcome.

Replication rates (when using the binary success criteria) ranged from 30% (16) to 78% (23) in large-scale social and behavioural science replication projects. There is ongoing discussion to determine the expected or desired rate (1,38–42), as it is well attested that not all claims are replicable. Validating the replicability of claims through direct replication, often by employing a much larger sample size than the original study, is expensive in terms of money, time, and resources (43,44). Having to replicate every finding in all literature in the social and behavioral sciences is impossible. Therefore, understanding which claims are likely to replicate (or not) without the need of performing a direct replication is highly valuable.

1.2 Using Human Forecasts to Evaluate Replicability

Initially, the outcomes of the Reproducibility Project: Psychology (21) were predicted by Dreber et al (45) using two crowd-sourcing methods, namely prediction markets and surveys. These methods proved effective as forecasts provided by human experts and aggregated through prediction markets could predict if a claim would be replicated with a 70% success rate. These methods were repeated for predicting outcomes of replication in Economics (22), Social Science (24) and other psychology projects (46), with success rates ranging from 61% to 86%. The forecasting of replication outcomes has a number of uses (45). Forecasting can be used to establish a consensus regarding a theory, and to gauge the academic community expectations, capturing how novel or surprising results are. This may be especially valuable for non-experts in field, or alongside reporting in media, so readers are able to interpret findings in context of the scientific consensus. Forecasting is also able to allocate replication resources more efficiently – by assigning a low priority to the replication of highly (or conversely lowly) rated papers, focusing on papers where the forecasters are most uncertain. Replication forecasting could also be used to set priors in Bayesian analysis, or weight studies in a meta-analysis. Forecasting could be used to assist editors when making publication decisions, either in the desk rejection stage, or to act as another reviewer.

The success of crowdsourcing comes from the principle of the wisdom of the crowd – where an aggregation of forecasters will often outperform a single forecaster. Despite individual members of the crowd having weak signals, errors are averaged out, resulting in a strong aggregated prediction (47,48). Prediction markets are designed to leverage the wisdom of the crowd by aggregating widely dispersed signals amongst a crowd of agents (49,50). The theory behind prediction markets is based on the efficient-market hypothesis, where the price of an asset reflects all available information and proper scoring, and agents are incentivised to report their true

beliefs. There is evidence that markets are effective at aggregating dispersed information (51,52) and have been used to predict election results in the Iowa Electronic Markets, providing empirical evidence that prediction markets can outperform other forecasting tools such as polls (53). These markets have also been used in the corporate world by popular firms such as Google and Ford to forecast their sales (54).

In prediction markets, agents trade assets with payoffs tied to the outcome of a future event. Using Arrow-Debreu securities, an asset pays \$1 if the event occurs and \$0 otherwise. Other types of assets can also be used (55). The market price of assets prove informative to the outcome of the event, with the price often being interpreted as the forecasted probability of an event occurring (56). Prediction markets can employ a ‘market maker’ that the agents trade with, thereby ensuring that an asset can always be bought or sold at a given price. Prediction markets distinguish themselves from other crowdsourced methods, such as surveys, as they elicit, aggregate and disseminate beliefs through a single index, namely, the market price (52). In the context of replication-based forecasting projects, surveys aggregated by averaging have never outperformed prediction markets.

1.3 SCORE: Systematizing Confidence in Open Research

Evidence

In response to the replication crisis in the social and behavioural sciences, and the success of the crowd-sourced replication forecasting projects, the Defense Advanced Research Projects Agency (DARPA) initiated a project to address the lack of confidence in scientific claims, named Systematizing Confidence in Open Research Evidence (SCORE) (57). The SCORE project was motivated by the Department of Defense (DoD) losing confidence in the research it leverages in its

operations (58). The SCORE project had 3 phases. In the first phase, 30,000 scientific papers were collected from 62 journals dating between 2009 and 2018. A random sample of 3000 papers was selected, and details about these papers were extracted, including a specific claim and the statistical evidence for this claim. In Phase 2, two teams of researchers sought to provide confidence scores for each of the 3000 claims. The confidence scores are a prediction of a claim's replicability, driven by human forecasts. A team from Melbourne University used the IDEA (Investigate, Discuss, Estimate, Aggregate) protocol to complete group evaluations of claims (59–61). In parallel, I, along with a large international collaboration (known as Replication Markets), also estimated confidence scores using a method adapted from previously used 'proof of concept' forecasting studies (22,24,45,46). This thesis includes research from my work with these teams.

Our method was centred around using surveys and prediction markets to assess the replicability of claims in 10 monthly rounds. Each month, incentivised surveys for 300 claims were completed by human forecasters, who answered a series of questions regarding the claims, including the estimated probability of finding a statistically significant effect in the same direction as the original study, if that study was to be replicated. Forecasters were unable to see the responses of any other forecasters to ensure independence. Once the survey was completed, a prediction market platform ran for two weeks, with forecasters being free to trade in any of individual prediction markets relating to the 300 claims provided for that month. This process created the opportunity for several confidence scores to be calculated for each claim, by aggregating the elicited forecasts in different ways. My role within this project was focused on the experimental design and the design and implementation of statistical analysis of the forecasts. The human forecasts were judged based on a small sample of the 3000 claims that were selected for direct replication. Phase 3, which I am not involved in, is based on automating the human forecasting component using machine learning. A full description of SCORE can be found in

Appendix 1, including an overview of the methodology used by all teams involved. The overarching aim of SCORE is to develop and test methods of assessing the credibility of scientific claims without direct replication – something that has only been attempted on a small scale in the past.

There were two unplanned additions to this research due to the COVID-19 pandemic. As a direct extension to the SCORE programme, and in addition to the 3000 social and behavioural claims, we also forecasted the replicability of 100 claims related to COVID-19. The same survey and prediction markets methods were applied to forecasting the COVID-19 claims. In addition, the research team gathered forecasts on whether or not a COVID-19-related pre-print would be published in a journal, the quality of the journal, and how many citations it will receive within a given timeframe.

The primary motivation behind forecasting replicability in the SCORE project is to provide guidance to practitioners, policy setters and decision makers as to the credibility of scientific claims in the social and behavioural sciences. Large organisations, both private and public³ often leverage social and behavioural science research to “design plans, guide investments, assess outcomes, and build models of human social systems and behaviors” (62). Therefore, non-replicating research has real world implications for such organisations who use the research. The assessments of credibility as included in this thesis can help to provide a level of confidence in research and assist with increasing the effectiveness of real-world uses for social and behavioural science.

³ Including the funders of the SCORE project: DARPA and the US Department of Defense

1.4 Thesis Structure

This thesis focuses on the statistical analysis of human forecasts of replication outcomes as part of the SCORE project, supplemented by additional some smaller ‘spin-off’ projects. Following this introduction (chapter 1) the structure of this thesis is as follows.

Chapter 2 contains a paper published in PLOS One under the title “*Predicting replicability - analysis of survey and prediction market data from large-scale forecasting projects*”. This paper presents a meta-analysis of four replication-based forecasting projects (22,24,45,46), analysing the degree to which experts in a given field can identify claims that are likely to replicate, as well as analysing methods used to elicit and aggregate human beliefs. As the four projects were very similar in methodology, the data of each could be combined to perform tests with much higher power than what was possible in the individual projects. Therefore, this paper provides stronger evidence as to the efficacy of the human forecasters to predict replicability. In addition, the increased sample size allows for new analyses. Much of the specifics of the SCORE methodology is informed by the findings from this paper. Specifically, the poor performance of aggregating the surveys via mean, the diminishing returns of accuracy of the prediction markets over time and the informativeness of p-values with respect to replication outcomes, all drove specific aspects of the methodology our team contributed to SCORE, as discussed in chapters 3, 4 and appendix 1. I am first author of this paper and was also the main contributor to both the analysis and report writing.

Chapter 3 includes the pre-registration documents for the experimental and statistical design for collecting, aggregating, and analysing forecasts of replicability collected for the DARPA SCORE project (for full programme details see appendix 1). These documents cover the initial meta round (as covered in chapter 4), the regular claim rounds and the final COVID-19 related round. The forecast collection as described in the pre-registration documents have already been

completed as the experiment ran from August 12, 2019, through to September 28, 2020. Due to COVID-19, the replications which can be used to validate our crowdsourced forecasts were not able to be conducted. As no results are available, no manuscript or report which presented results could be completed. Therefore, I include the pre-registration documents to represent the research that was completed by myself and the Replication Markets team.

Chapter 4 contains a paper published in Royal Society Open Science under the title “*Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme*”. This research focuses on the first round of surveys and prediction markets in which we sought ‘meta’ forecasts from the community. The forecasts do not relate to any single claim or finding but rather predict replicability across academic fields and across years of publications. This paper found that human forecasters expect there to be differences in replication rates among fields. The scope for SCORE is broad – much broader than past replication projects which focus on a single field. This broad scope includes fields where no attempt to quantify credibility through large scale replications have been undertaken. In these fields, forecasts present an indication of field wide reproducibility without any ground truth replications. I am first author of this paper and the findings presented were used directly in calculating a set of confidence scores as described in chapter 3.

Chapters 5 and 6 contain the papers; “*Creative destruction in science*” (published in *Organizational Behavior and Human Decision Processes*) and “*A creative destruction approach to replication: Implicit work and sex morality across cultures*” (published in *Journal of Experimental Social Psychology*) respectively. These two papers use replication to test multiple hypothesis in the same subject space, to revise and update theories. In addition to testing the hypotheses, these projects also used human experts to forecast the outcomes of the tests. “*Creative destruction in science*” (chapter 6) focuses on theories relating to gender and hiring. We found that forecasters can predict the direction and size of replication effects, including simple, interaction and moderator

effects. We also found forecasters' political beliefs regarding gender and hiring practices did not affect their accuracy. "*A creative destruction approach to replication: Implicit work and sex morality across cultures*" (chapter 7) focused on culture and work theories. We found that human forecasts did correlate with realised effect sizes but were appreciatively more accurate in estimating direction and relative size of effect sizes compared with absolute effect sizes. My role in these papers included experimental and statistical design of forecasting components including drafting the pre-analysis plan, statistical analysis, report writing and editing of the forecasting components. As my contribution was limited to one part of a much larger project (i.e the forecasting) I have only included the sections of the publication relevant to the forecasting. In both papers this includes one section in the main paper and more detailed analyses in supplementals, including the pre-analysis plans.

Chapter 7 is the paper "*Forecasting the citation and publishing outcomes of COVID-19 pre-prints*". This study sought to forecast the publication outcomes and future citations of COVID-19 related preprints. We forecast whether 400 preprints will not be published, published in a low or medium impact journal, or published in a high impact journal after one year. We also forecast the citation rank relative to other preprints in the study. We find that forecasters can predict the publication outcomes of the preprints, however they are much more accurate at predicting citation ranks, potentially due to the less stochastic nature of citations as compared to publications. I am first author and the main contributor to experimental and statistical design, manuscript drafting and editing.

Chapter 8 concludes and discusses the research found in this thesis. I provide a synthesis of the findings in this thesis, followed by a description of the contribution to knowledge. I finish with areas of further study and brief final thoughts.

1.5 References

1. Baker M. 1,500 scientists lift the lid on reproducibility. *Nat News*. 2016 May 26;533(7604):452.
2. Ioannidis J. Why Most Published Research Findings Are False. *PLOS Med*. 2005 Aug 30;2(8):e124.
3. Begley CG, Ellis LM. Raise standards for preclinical cancer research. *Nature*. 2012 Mar;483(7391):531–3.
4. Ioannidis JPA. Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *JAMA*. 2005 Jul 13;294(2):218–28.
5. Mobley A, Linder SK, Braeuer R, Ellis LM, Zwelling L. A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic. *PLoS ONE*. 2013 May 15 [cited 2021 Mar 24];8(5). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3655010/>
6. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov*. 2011 Sep;10(9):712–712.
7. Casadevall A, Fang FC. Reproducible Science. *Infect Immun*. 2010 Dec 1;78(12):4972–5.
8. Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA. An open investigation of the reproducibility of cancer biology research. Rodgers P, editor. *eLife*. 2014 Dec 10;3:e04333.
9. Morrison SJ. Time to do something about reproducibility. *eLife*. 2014 Dec 10;3:e03981.
10. Schloss PD. Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. *mBio* 2018 Jul 5 [cited 2021 Apr 14];9(3). Available from: <https://mbio-asm-org.ezproxy.massey.ac.nz/content/9/3/e00525-18>
11. Drummond DC. Replicability is not Reproducibility: Nor is it Good Science. In 2009 [cited 2020 Feb 8]. Available from: <http://cogprints.org/7691/>
12. Gundersen OE, Gil Y, Aha DW. On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. *AI Mag*. 2018 Sep 28;39(3):56–68.
13. Gundersen OE, Kjensmo S. State of the Art: Reproducibility in Artificial Intelligence. In: *Thirty-Second AAAI Conference on Artificial Intelligence 2018* [cited 2020 Feb 19]. Available from: <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17248>
14. Hunter JE. The Desperate Need for Replications. *J Consum Res*. 2001 Jun 1;28(1):149–58.
15. Halperin I, Vigotsky AD, Foster C, Pyne D. Strengthening the Practice of Exercise and Sport-Science Research. *Int J Sports Physiol Perform*. 2018;

16. Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, Banks JB, et al. Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J Exp Soc Psychol*. 2016 Nov 1;67:68–82.
17. Ebersole CR, Mathur MB, Baranski E, Bart-Plange D-J, Buttrick NR, Chartier CR, et al. Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. *Adv Methods Pract Psychol Sci*. 2020 Sep 1;3(3):309–31.
18. Klein RA, Ratliff KA, Vianello M, Adams Jr. RB, Bahník Š, Bernstein MJ, et al. Investigating variation in replicability: A “many labs” replication project. *Soc Psychol*. 2014;45(3):142–52.
19. Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB, Alper S, et al. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Adv Methods Pract Psychol Sci*. 2018 Dec;1(4):443–90.
20. Klein RA, Cook CL, Ebersole CR, Vitiello C, Nosek BA, Chartier CR, et al. Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement *PsyArXiv*; 2019 [cited 2021 Mar 24]. Available from: <https://psyarxiv.com/vef2c/>
21. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015 Aug 28;349(6251):aac4716.
22. Camerer CF, Dreber A, Forsell E, Ho T-H, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. *Science*. 2016 Mar 25;351(6280):1433–6.
23. Cova F, Strickland B, Abatista A, Allard A, Andow J, Attie M, et al. Estimating the Reproducibility of Experimental Philosophy. *Rev Philos Psychol*. 2021 Mar 1;12(1):9–44.
24. Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat Hum Behav*. 2018 Sep;2(9):637–44.
25. Fanelli D. “Positive” Results Increase Down the Hierarchy of the Sciences. *PLOS ONE*. 2010 Jul 4;5(4):e10068.
26. Fanelli D. Negative results are disappearing from most disciplines and countries. *Scientometrics*. 2012 Mar 1;90(3):891–904.
27. John LK, Loewenstein G, Prelec D. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychol Sci*. 2012 May 1;23(5):524–32.
28. Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol Sci*. 2011 Nov 1;22(11):1359–66.
29. Fidler F, Wilcox J. Reproducibility of Scientific Results. In: Zalta EN, editor. *The Stanford Encyclopedia of Philosophy*. Winter 2018. Metaphysics Research Lab, Stanford University; 2018 [cited 2019 Nov 14]. Available from: <https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility/>

30. Schmidt S. Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Rev Gen Psychol.* 2009 Jun 1;13(2):90–100.
31. Shadish WR, Cook TD, Campbell DT. *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin,; 2002.
32. Nosek BA, Errington TM. What is replication? *PLOS Biol.* 2020 Mar 27;18(3):e3000691.
33. Artner R, Verliefdede T, Steegen S, Gomes S, Traets F, Tuerlinckx F, et al. The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychol Methods.* 2020.
34. Crandall CS, Sherman JW. On the scientific superiority of conceptual replications for scientific progress. *J Exp Soc Psychol.* 2016 Sep 1;66:93–9.
35. Tierney W, Hardy JH, Ebersole CR, Leavitt K, Viganola D, Clemente EG, et al. Creative destruction in science. *Organ Behav Hum Decis Process.* 2020 Nov 1;161:291–309.
36. Tierney W, Hardy J, Ebersole CR, Viganola D, Clemente EG, Gordon M, et al. A creative destruction approach to replication: Implicit work and sex morality across cultures. *J Exp Soc Psychol.* 2021 Mar 1;93:104060.
37. Gordon M, Viganola D, Dreber A, Johannesson M, Pfeiffer T. Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. *PLOS ONE.* 2021 Apr 14;16(4):e0248780.
38. Christensen G, Miguel E. Transparency, Reproducibility, and the Credibility of Economics Research. *J Econ Lit.* 2018 Sep;56(3):920–80.
39. Etz A, Vandekerckhove J. A Bayesian Perspective on the Reproducibility Project: Psychology. *PloS One.* 2016;11(2):e0149794.
40. Fanelli D. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proc Natl Acad Sci.* 2018 Mar 13;115(11):2628–31.
41. Nosek BA, Hardwicke TE, Moshontz H, Allard A, Corker KS, Almenberg AD, et al. Replicability, Robustness, and Reproducibility in Psychological Science. *PsyArXiv*; 2021 [cited 2021 Mar 25]. Available from: <https://psyarxiv.com/ksfvq/>
42. Pashler H, Harris CR. Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspect Psychol Sci.* 2012 Nov 1;7(6):531–6.
43. Coles NA, Tiokhin L, Scheel AM, Isager PM, Lakens D. The Costs and Benefits of Replication Studies. *PsyArXiv*; 2018 Jan [cited 2020 Jan 20]. Available from: <https://osf.io/c8akj>
44. Isager PM, Aert RCM van, Bahník Š, Brandt M, DeSoto KA, Giner-Sorolla R, et al. Deciding what to replicate: A formal definition of “replication value” and a decision model for replication study selection. *MetaArXiv*; 2020 [cited 2021 Mar 22]. Available from: <https://osf.io/preprints/metaarxiv/2gurz/>

45. Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, et al. Using prediction markets to estimate the reproducibility of scientific research. *Proc Natl Acad Sci*. 2015 Dec 15;112(50):15343–7.
46. Forsell E, Viganola D, Pfeiffer T, Almenberg J, Wilson B, Chen Y, et al. Predicting replication outcomes in the Many Labs 2 study. *J Econ Psychol*. 2019 Dec 1;75:102117.
47. Yi SKM, Steyvers M, Lee MD, Dry MJ. The Wisdom of the Crowd in Combinatorial Problems. *Cogn Sci*. 2012;36(3):452–70.
48. Surowiecki J. *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York, NY, US: Doubleday & Co; 2004. xxi, 296 p. (The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations).
49. Arrow KJ, Forsythe R, Gorham M, Hahn R, Hanson R, Ledyard JO, et al. The Promise of Prediction Markets. *Science*. 2008 May 16;320(5878):877–8.
50. Wolfers J, Zitzewitz E. *Prediction Markets in Theory and Practice*. National Bureau of Economic Research; 2006 Mar [cited 2019 Oct 16]. Report No.: 12083. Available from: <http://www.nber.org/papers/w12083>
51. Plott CR, Chen K-Y. *Information Aggregation Mechanisms: Concept, Design and Implementation for a Sales Forecasting Problem*, Pasadena, CA: California Institute of Technology; 2002 [cited 2021 Apr 12]. Available from: <https://resolver.caltech.edu/CaltechAUTHORS:20140317-135547085>
52. Plott CR, Sunder S. Rational Expectations and the Aggregation of Diverse Information in Laboratory Security Markets. *Econometrica*. 1988;56(5):1085–118.
53. Berg JE, Nelson FD, Rietz TA. Prediction market accuracy in the long run. *Int J Forecast*. 2008 Apr 1;24(2):285–300.
54. Cowgill B, Zitzewitz E. Corporate Prediction Markets: Evidence from Google, Ford, and Firm X *. *Rev Econ Stud*. 2015 Oct 1;82(4):1309–41.
55. Wolfers J, Zitzewitz E. Prediction markets. *J Econ Perspect*. 2004;18(2):107–26.
56. Wolfers J, Zitzewitz E. Interpreting Prediction Market Prices as Probabilities. National Bureau of Economic Research; 2006 May [cited 2019 Aug 5]. Report No.: 12200. Available from: <http://www.nber.org/papers/w12200>
57. Gordon M, Viganola D, Bishop M, Chen Y, Dreber A, Goldfedder B, et al. Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *R Soc Open Sci*. 2020 Jul 22;7(7):200566.
58. Defense Sciences Office. *Systematizing Confidence in Open Research and Evidence (SCORE)*. DARPADSO Broad Agency Announc HR001118S0047. 2018 Jun 12; Available from: <https://research-vp.tau.ac.il/sites/resauth.tau.ac.il/files/DARPA-SCORE-DSO-2018.pdf>

59. Pearson R, Fraser H, Bush M, Mody F, Widjaja I, Head A, et al. Eliciting group judgements about replicability: a technical implementation of the IDEA Protocol. 2021 [cited 2021 Sep 23]. Available from: <http://scholarspace.manoa.hawaii.edu/handle/10125/70666>
60. Fraser H, Bush M, Wintle B, Mody F, Smith E, Hanea A, et al. Predicting reliability through structured expert elicitation with repliCATS (Collaborative Assessments for Trustworthy Science). MetaArXiv; 2021 [cited 2021 Sep 23]. Available from: <https://osf.io/preprints/metaarxiv/2pczv/>
61. Hanea A, Wilkinson DP, McBride M, Lyon A, Ravenzwaaij D van, Thorn FS, et al. Mathematically aggregating experts' predictions of possible futures. MetaArXiv; 2021 [cited 2021 Sep 23]. Available from: <https://osf.io/preprints/metaarxiv/rxmh7/>
62. Witkop G. Systematizing Confidence in Open Research and Evidence. [cited 2021 Dec 15]. Available from: <https://www.darpa.mil/program/systematizing-confidence-in-open-research-and-evidence>

2 Chapter 2: Predicting replicability - analysis of survey and prediction market data from large-scale forecasting projects

This chapter consists of the paper “*Predicting replicability - analysis of survey and prediction market data from large-scale forecasting projects*” that was published in 2021 in the journal “*PLOS ONE*”. This paper serves both as a meta-analysis of previous replication forecasting projects to confirm their findings and to use the higher power made available by pooling data to derive new findings and insights. The findings from this paper heavily informed the subsequent replication forecasting projects included in this thesis. This chapter also serves as a literature review as it takes an in depth look at all the previous replication forecasting literature. My role for this paper included performing the statistical design, conducting the statistical analysis and wrote the manuscript.

The reference for this paper is: Gordon M, Viganola D, Dreber A, Johannesson M, Pfeiffer T.

Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. PLOS ONE. 2021;16: e0248780.

To align with the formatting and referencing style of this thesis, there are some changes in formatting and referencing style of the published paper

Predicting replicability - analysis of survey and prediction market data from large-scale forecasting projects

2.1 Abstract

The reproducibility of published research has become an important topic in science policy. A number of large-scale replication projects have been conducted to gauge the overall reproducibility in specific academic fields. Here, we present an analysis of data from four studies which sought to forecast the outcomes of replication projects in the social and behavioural sciences, using human experts who participated in prediction markets and answered surveys. Because the number of findings replicated and predicted in each individual study was small, pooling the data offers an opportunity to evaluate hypotheses regarding the performance of prediction markets and surveys at a higher power. In total, peer beliefs were elicited for the replication outcomes of 103 published findings. We find there is information within the scientific community about the replicability of scientific findings, and that both surveys and prediction markets can be used to elicit and aggregate this information. Our results show prediction markets can determine the outcomes of direct replications with 73% accuracy ($n=103$). Both the prediction market prices, and the average survey responses are correlated with outcomes (0.581 and 0.564 respectively, both $p < .001$). We also found a significant relationship between p-values of the original findings and replication outcomes. The dataset is made available through the R package “pooledmaRket” and can be used to

further study community beliefs towards replications outcomes as elicited in the surveys and prediction markets.

2.2 Introduction

The communication of research findings in scientific publications plays a crucial role in the practice of science. However, relatively little is known about how reliable and representative the disseminated pieces of information are (1,2). Concerns have been raised about the credibility of published results following John Ioannidis' landmark essay, "Why most published findings are false" (3), and the identification of a considerable number of studies that turned out to be false positives (4,5). In response, several large-scale replication projects were initiated in the fields of psychology, experimental economics, and the social sciences more generally (6–14) to systematically evaluate a large sample of findings through direct replication. The rate of successful replication (commonly defined as a result with a statistically significant effect size in the same direction as the original effect) in these projects ranges from 36% to 62%. This rate, however, cannot be easily compared across projects because key features such as the inclusion criteria and replication power differed across projects. For a discussion of these findings in the context of the 'replication crisis' see refs (1,15–18).

Four of the large-scale replications projects were accompanied by forecasting projects: the Reproducibility Project: Psychology (RPP) (12); the Experimental Economics Replication Project (EERP) (6); the Many Labs 2 Project (ML2) (10); and the Social Science Replication Project (SSRP) (7). The replications and the forecasting results were included in a single publication for EERP(6) and SSRP(7). For RPP (12) and ML2 (10), the forecasting studies appeared separately(19,20). In each of the replication projects, a set of original findings published in the scientific literature were selected to be repeated via direct replication on new participants and

typically larger sample sizes. The purpose of the associated prediction market studies was to investigate whether information elicited from within research communities can be used to predict which findings in the replication projects are likely to replicate; and whether prediction markets and surveys are useful mechanisms for eliciting such information from scientists. The previously published results of the forecasting studies show that the research community can predict which findings are likelihood to replicate – with varying degrees of accuracy. In total, peer beliefs were elicited for the replication outcomes of 103 published findings in the social and behavioural sciences. We have made the resulting dataset available in an R package – ‘pooledmaRket’.

In this paper, we present an analysis of this pooled dataset, which allows for both testing hypotheses with substantially higher statistical power and for conducting additional analyses not possible in the previous smaller studies. In the following, we provide a methods section with a brief review of the methodology used in the large-scale replication projects and the prediction market studies as well as how the dataset is analysed in this paper. This is followed by the results of our statistical analysis and a discussion.

2.3 Methods

2.3.1 Replication Projects

Within the replication projects (6,7,10,12), original findings published in the scientific literature were selected based on a set of pre-defined criteria, including research methodology, specific target journals, and time windows. Typically, one key finding of a publication was selected to be replicated with a methodology as close as possible to the original paper. The authors of the original papers were contacted and asked to provide feedback on the replication designs before starting the data collection for the replications.

For RPP, EERP, and SSRP, a replication was deemed successful if it found a ‘significant effect size at 5% in the same direction of the original study’ (12,21); for ML2, a replication was deemed successful if it found ‘a significant effect size in the same direction of the original study and a p-value smaller than 0.0001’ (10). The latter definition of a successful replication is more stringent because the power of the replications in the ML2 project is higher with multiple laboratories conducting replications. The large-scale replication projects also report additional replication outcomes such as effect sizes.

Statistical power for the replications was typically higher than for the original findings. RPP and EERP had a statistical power of about 90% to find the original effect size. The power was increased substantially for the SSRP project following concerns that effect sizes for original findings may be inflated (12,22), which increases the chances of false negatives among the replication outcomes in the RPP and EERP projects. This was done by using a 2-stage design, where sample sizes in the first stage were such that replications had 90% power to detect 75% of the original effect size. The second stage was conducted if the first stage found a negative result, and together the samples of the two stages led to the replications having 90% power to detect 50% of the original effect size. This two-stage approach is further explained below. In the ML2 study, replications were conducted at multiple sites with large sample sizes, resulting in a substantially higher power.

2.3.2 Forecasting Studies

The four forecasting studies associated with the replication projects investigated the extent to which prediction markets and surveys can be used to forecast the replicability of published findings. Before the replication outcomes became public information, peer researchers participated

in a survey eliciting beliefs about the replication probability for findings within the replication projects and thereafter participated in prediction markets.

In the prediction markets, participants were endowed with tokens that could be used to buy and sell contracts each of which paid one token if a finding was replicated, and zero tokens if it was not replicated. At the end of the study, tokens earned from the contracts were converted to monetary rewards. The emerging price for the contracts traded in the market can be interpreted, with some caveats (23), as a collective forecast of the probability of a study replicating. An automated market maker implementing a logarithmic market scoring rule was used to determine prices (24). The prediction markets were open for two weeks in RPP, ML2, and SSRP, and for 10 days in EERP. The most relevant information for forecasting, including the power of the replications, was embedded in the survey and in the market questions, and the links to the original publications were provided. The forecasters were also provided with the pre-replication versions of the replication reports detailing the design and planned analyses of each replication. In the case of ML2, where many replications were performed for each finding, overall protocols on the replications were provided in lieu of specific replication reports.

Participants were recruited via blogs, mailing lists, and Twitter – with the focus on people working within academia. Some participants who filled out the survey did not participate in the prediction markets. The data presented here is restricted to only those participants who actively participated in the markets, therefore a participant had to trade in at least one market to be included in the survey data. As a result, both the survey and prediction market data are based on the same participants.

Study procedures for each of the forecasting studies meet the guidelines of written documentation provided by the Ethical Review Board in Sweden. Since no sensitive personal

information was collected, we did not consider the study being eligible for review or that it required documenting informed consent in line with the Swedish legislation on ethical review. We did not obtain formal ethics review for the forecasting of RPP, EERP and ML2 outcomes, and did not explicitly document informed consent. Participants who were invited to participate in these studies received an information sheet about the design and purpose of the study and subsequently could choose whether to contribute to the forecasting. For SSRP, an Ethical Review Board application was submitted to The Regional Ethical Review Board in Stockholm (2016/1063-31/5). The Ethical Review Board of Stockholm decided that an application was not required by Swedish legislation and that they therefore offered only an "advisory opinion" in which they noted that they had no ethical objections against the study. Informed consent was explicitly documented for SSRP.

The following subsections provide study specific details; further information is available in the original publications.

2.3.2.1 RPP

The forecasting study by Dreber et al. (19) was done in conjunction with the Reproducibility Project: Psychology (12). In RPP, a sample of findings published in the *Journal of Personality and Social Psychology*, *Psychological Science*, and *Journal of Experimental Psychology* was replicated. The overall replication rate was 36%. The total RPP included 97 original findings, 44 of which were included in both prediction markets and surveys. Dreber et al. ran these 44 prediction markets and 44 surveys in two separate batches in November 2012 and in October 2014 to study whether researchers' beliefs carry useful information about the probability of successful replication. For these 44 studies, 41 replications had been finished at the time of publication. One finding is excluded as it does not have relevant survey forecasts, leaving a total of 40 sets of forecasts included in this dataset. Of the 40 findings included here, prediction markets

correctly predicted the outcome of the replications 70% of the time, compared with 58% for the survey. The overall replication rate of the included 40 findings was 37.5% (see Table 1).

2.3.2.2 EERP

Camerer et al. (6) replicated 18 findings in the field of experimental economics, published in two of the top economic journals (American Economic Review and Quarterly Journal of Economics). The process for selecting the finding to be replicated from a publication was as follows: (1) select the most central finding in the paper (among the between-subject treatment comparisons) based on to what extent the findings were emphasized in the published versions; (2) if there was more than one equally central finding, the finding (if any) related to efficiency was picked, as efficiency is central to economics; (3) if several findings remained and they were from different separate experiments, the last experiment (in line with RPP) was chosen; (4) if several findings still remained, one of those findings was randomly selected for the replication. A fraction of 61% of replications were successful. Both the markets and the survey correctly categorized 11 findings out of 18 (61%).

2.3.2.3 ML2

Forsell et al. (25) presents forecasts for replications included in the ML2 project (10), a further large-scale replication project led by the Open Science Collaboration. One of the aims of the ML2 project was to guarantee high-quality standards for the replications of classic and contemporary findings in psychology by using large sample sizes across different cultures and laboratories and requiring replication protocols to be peer-reviewed in advance. The findings were selected by the authors of the ML2 project, with the aim of assuring diversity and plurality of findings. The realized replication rate for the ML2 project was 46% (11 successful replications out of 24 findings analysed). Although ML2 replicated in total 28 findings, replication outcomes were

only forecasted for 24 of these. The excluded findings focused on cultural differences in effect sizes across different samples. Note that when including all 28 findings, the replication rate of the Many Labs 2 project (10) increases to 50% (14/28). Further detail is given in Appendix A in Forsell et al. (20). The prediction markets correctly predicted 75% of the replication outcomes. As a comparison, the survey correctly predicted 67% of replication outcomes.

2.3.2.4 SSRP

SSRP is a replication project covering 21 experimental social science studies published in two high-impact interdisciplinary journals, *Science* and *Nature* (7). SSRP was specifically designed to address the issue of inflated effect sizes in original findings. There were three criteria for selecting findings within publications (presented in descending order): (1) select the first finding that reports a significant effect; (2) select the statistically significant finding identified as the most important; (3) randomly select a single finding in cases of more than one equally central result. In line with previous studies, Camerer et al (7), also ran prediction markets and prediction surveys to forecast whether the selected studies will replicate. The design of SSRP for conducting replications differed from the previous projects in that it was structured in two stages: the first stage provided 90% power to detect 75% of the original effect size; if the replication failed, stage 2 started, and the data collection continued until there was 90% power of detecting 50% of the original effect size (pooling data from the stage 1 and stage 2 collection phases). Based on all the data collected, 62% of the 21 findings were successfully replicated. The prediction markets followed a similar structure of the data collection: participants were randomized in two groups: in treatment 1, beliefs about replicability in stage 1 were elicited; in treatment 2, beliefs about replicability in both stage 1 and stage 2 were elicited. In this paper, we report the results about treatment 2 only, as the replication results after Stage 2 are the most informative about the replication outcome.

2.3.3 Pooled Dataset

Due to the high similarity in research topic and design of the four forecasting studies, they can be pooled into a single dataset. The pooled data can be downloaded within the R package which can be accessed at <https://github.com/MichaelbGordon/pooledmaRket>. The dataset is presented in three separate tables, combined from the four forecasting studies as well as a codebook which provides details on each of the columns within each dataset. Each table represents the key parts of the studies; replication outcomes and original findings features, survey responses, and prediction market trades. Each of these tables is made available in the R package, as well as example code of aggregation methods.

In order to analyse the performance of the prediction markets we typically take the market price at time of closing as the aggregated prediction of the market. For the survey we aggregate primarily with simple mean, but also provide performance of several other aggregations. In total we analyse data from over 15,000 forecasts across the 103 findings, made up of 7850 trades and 7380 survey responses.

2.4 Results

In this section, we report and comment on the outcomes of the descriptive and statistical analyses performed to compare the prediction markets results and the survey results. For all the results reported below, the tests are interpreted as two-tailed tests and a p-value < 0.005 is interpreted as “statistically significant” while a p-value < 0.05 as “suggestive” evidence, in line with the recommendation of Benjamin et al. (26).

2.4.1 Observed and Forecasted Replication Rates

Successful rates of replication ranged from 38% to 62%, with an overall rate of 49% for the 103 findings included in the dataset (Table 1). The Replication Project Psychology had the lowest overall replication rate of 38%. Many Labs 2 had the second-lowest replication rate with 11 out of 24 studies successfully replicating (45.8%). The Experimental Economics Replication Project and Social Studies Replication Project both have replication rates around 60%.

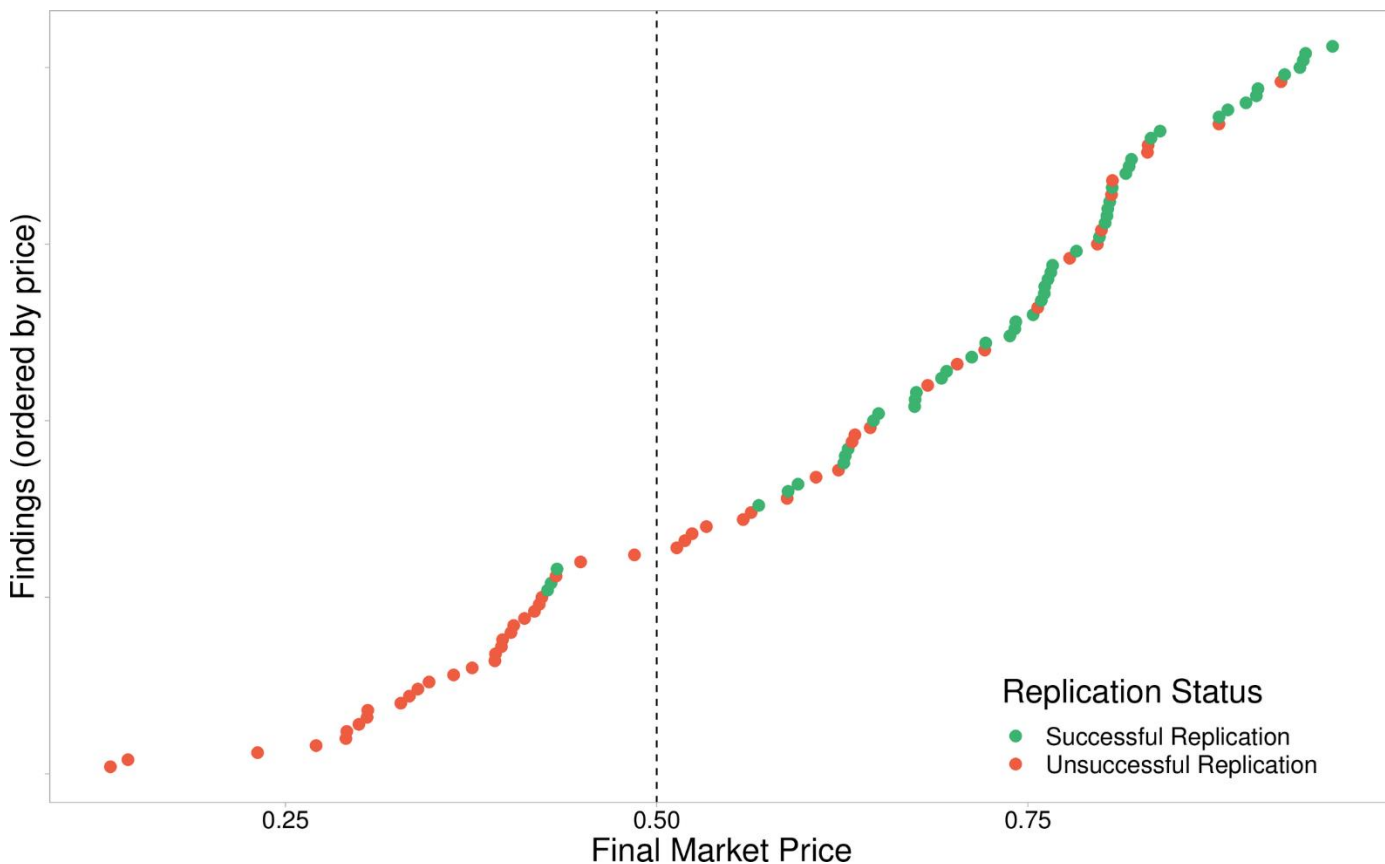
Table 2-1: Main features of individual projects. This table contains key characteristics and summaries of the datasets and the pooled data. In calculations of correct forecasts by the prediction market and survey, we interpreted a final price of 0.50 or greater as prediction of a successful replication; if the final price is lower than 0.50, we interpret this as prediction of a failed replication. Overall, the actual replication rate was 49%, indicating that the forecasters were overconfident with the average market price being 0.627. Prediction Markets tend to outperform surveys when forecasting replication success when considering overall accuracy – 73% compared to 66%.

	RPP	EERP	ML2	SSRP	Pooled data
Replication Study	Ref (12)	Ref (6)	Ref (10)	Ref (7)	
Forecasting Study	Ref (19)		Ref (25)		
Field of study	Experimental Psychology	Experimental Economics	Experimental Psychology	Experimental Social Science	
Source Journals	JPSP, PS, JEP (2008)	AER, QJE (2011-2014)	Several psychology outlets, including JEP, JPSP, PS (1977-2014)	Science, Nature (2010-2015)	
Replicated Findings	40	18	24	21	103
Successful replications	15 (37.5%)	11 (61.1%)	11 (45.8%)	13 (61.9%)	51 (49%)
Mean beliefs - Prediction Market	0.556	0.751	0.644	0.634	0.627
Correct – Prediction Markets (%)	28(70%)	11 (61%)	18 (75%)	18 (86%)	76 (73%)
Mean Absolute Error – Prediction Market	0.431	0.414	0.354	0.303	0.384

Mean beliefs - survey	0.546	0.711	0.647	0.605	0.610
Correct - Survey (%)	23 (58%)	11 (61%)	16 (67%)	18 (86%)	68 (66%)
Spearman Correlation – Prediction Market and Survey beliefs	0.736	0.792	0.947	0.845	0.837
Spearman Correlation – Replication Outcomes and Prediction Market	0.418	0.297	0.755	0.842	0.568
Spearman Correlation – Replication Outcomes and Survey beliefs	0.243	0.516	0.731	0.760	0.557

Expected replication rates as found within prediction markets range from 56% (for RPP) to 75% (for EERP). Surveys predicted replication rates from 55% (RPP) to 71% (EERP). Overall, the replication rates as expected by the community is around 60%. When comparing actual with expected replication rates, both the average survey ($M = 0.61$, $SD = 0.14$) and final market price ($M=0.63$, $SD=0.21$) tend to overestimate the actual rate of replication success ($M = 0.49$). Paired t tests found statistically significant difference between actual replication rate and the survey ($t(102) = -2.89$, $p = 0.0046$) and the market ($t(102) = -3.43$, $p = 0.00088$).

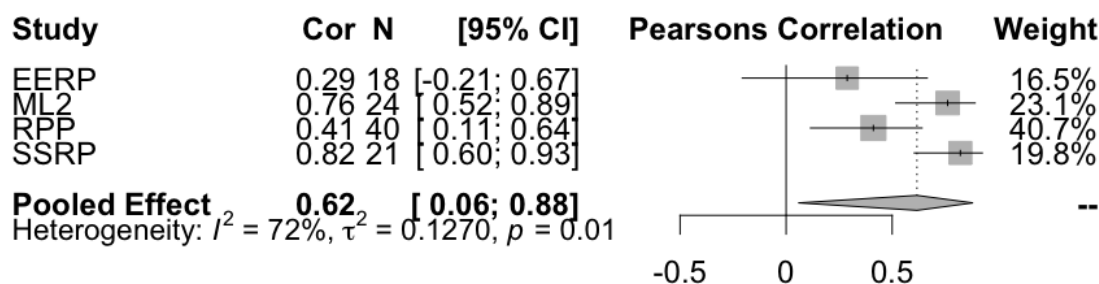
Figure 2-1. Market Beliefs. This figure plots the final prices of the 103 markets included within this dataset ordered by price. The green dots represent successful replications, and the non-replications are represented by the red dots. The vertical line at 0.5 indicates the binary cut off used to determine the markets aggregated prediction.



The binary outcome variable is correlated with both the survey responses ($r(101) = .564$, $p < .001$, 95% CI [0.42,0.68]) and market prices ($r(101) = .581$, $p < .001$, 95% CI [0.44,0.70]), as shown in Figure 1. When combining the final market price correlations in a random effects meta-analysis (using the conservative Hartung-Knapp-Sidik-Jonkman method to account for the small number of studies (27,28)), we find a pooled correlation of 0.62 ($p = 0.04$, 95% CI [0.05, 0.88]) (Figure 2). There is evidence of between study heterogeneity ($\tau^2 = 0.127$, 95% CI [0.007,2.41]) and $I^2 = 72\%$, 95% CI [20.4%,90.1%]). However, with the small number of studies included here, I^2

should be interpreted in context of its 95% CI⁴. All studies have positive correlations between final market prices and replication outcomes.

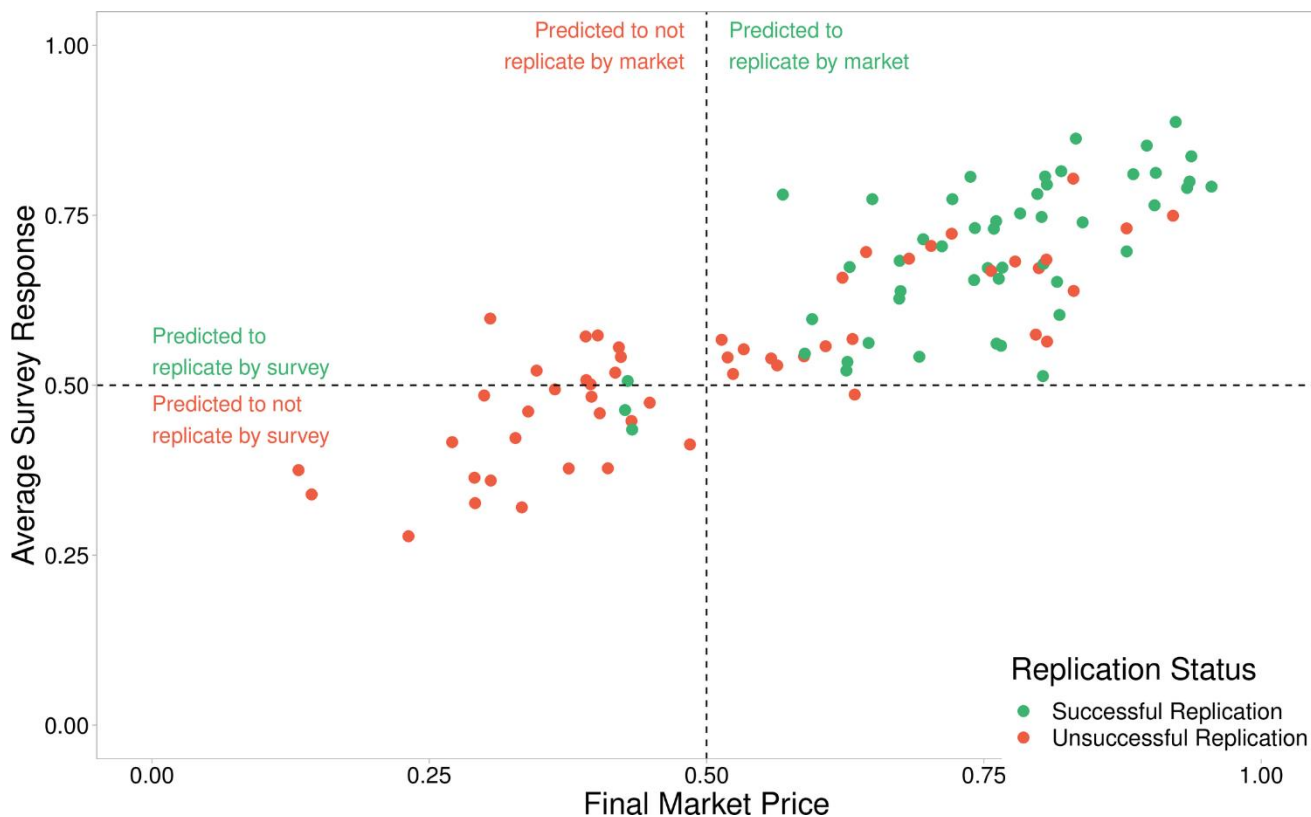
Figure 2-2. Market Price and Replication Outcome Correlation and Meta-Analysis. Final market price and replication outcome are shown to be correlated in each of the forecasting studies individually. Pooling the correlation using a random effects meta-analysis provides a pooled correlation of 0.62. There is some evidence for heterogeneity between studies with $I^2 = 72%$ (95% CI [20.4%,90.1%]), $\tau^2 = 0.127$ (95% CI [0.007,2.41]) and significant Q statistic ($Q(3) = 10.7, p=0.013$).



Market based and survey based forecasts in all four studies are highly correlated (RPP - $r_s(38) = 0.736, p < .001$; EERP - $r_s(16) = 0.792, p < .001$; SSRP - $r_s(19) = 0.845, p < .001$; ML2 - $r_s(22) = 0.947, p < .001$). When considering combined data the same high correlation is found ($r_s(101) = .837, p < .001$; $r(101) = .853, p < .001, 95\% \text{ CI } [0.79,0.90]$); as it emerges distinctly from Figure 2-3.

⁴ With limited number of forecasting projects ($n=4$), the estimation of I^2 will be noisy. Therefore, readers should analyse the range of I^2 as given by the confidence intervals rather than a point estimate.

Figure 2-3. Market and Survey Correlations. Final Market Prices and average survey responses are highly correlated ($r_s(101) = .837, p < .001$; $r(101) = .853, p < .001$). The dotted horizontal and vertical lines indicate the 0.5 cut off points used when applying a binary forecasting approach. The top left quadrant represents those findings which are predicted to replicate by survey but predicted to not replicate by market. The top right and bottom left quadrants contain findings where the markets and surveys agree, predicted to replicate and to not replicate respectively. The bottom right quadrant with a single finding, is where the study is predicted to replicate by the market but not by the survey. The colours of the findings show the replication outcome, with green indicating a successful replication outcome, and red indicating unsuccessful replication.



2.4.2 Accuracy of forecasts

In order to assess the effectiveness of forecasters providing beliefs via prediction markets and surveys, we analyse error rates for each method, and overall accuracy when adopting a binary approach. For the binary approach we interpret a final price of 0.50 or greater as prediction of a successful replication and a final price lower than 0.50 as prediction of a failed replication. The same rules applied for surveys: we computed the mean beliefs for each study and then interpreted that the survey predicts a successful replication if the average beliefs exceed 0.50 and a failed replication otherwise. Using this approach, the surveys never outperformed the markets. In two

cases (EERP and SSRP) they correctly categorize the same number of findings in the replicates/non-replicates dichotomy; in the other two studies, the markets do better (71% vs 58% in the RPP; 75% vs 67% in the ML2). Overall, the prediction markets had an accuracy of 73% (75 out of 103 studies), while the surveys had an accuracy of 66% (68 out of 103 studies). However, based on a chi-square test this difference is not statistically significant ($X^2(1) = 1.12, p = 0.29$).

Findings that do not replicate tend to have prediction market prices below the 0.50 threshold, while studies that do replicate are more concentrated above the 0.5 threshold. Out of the 31 findings that are predicted by the market not to replicate, only three eventually replicated, thus for these findings the market is correct more than 90% of the time. Alternatively, 25 of the 73 (66%) findings that were predicted to replicate did not. The survey-based predictions follow a similar pattern; of the 22 findings that are predicted to not replicate by the survey, only two eventually replicated and of the 81 studies that are predicted to successfully replicate, 33 did not replicate. Both the market and survey-based forecasts are more accurate when concluding that a study will not replicate rather than when concluding that a study will replicate, markets ($X^2(1) = 6.68, p = 0.01$; $X^2(1) = 4.45, p = 0.035$ for markets and surveys respectively). This may at least be partially due to the limited power of the replications in RPP and EERP, as some of the failed replications may be false negatives.

The absolute error, defined as the absolute difference between actual replication outcome (either 0 or 1) and the forecasted chance of replication, is used as an accuracy measure that does not entail a loss of information from binarizing the aggregated forecasts. The forecasts for SSRP had the lowest mean absolute error of 0.303 and 0.348 for the prediction markets and survey respectively. This was followed by ML2 (market error of 0.354 and survey error of 0.394) and RPP (market error of 0.431 and survey error of 0.485). Only in EERP there was a lower absolute mean error for the survey – 0.409 compared with 0.414 for the market. Across all 103 findings, the

absolute error of the prediction markets ($M = 0.384$) is significantly lower than the survey ($M = 0.423$) ($t(102) = 3.68$, $p < .001$). Prediction markets tend to provide more extreme forecasts with the final price ranges being larger in all four projects than the survey beliefs range. Quantifying extremeness as distance of a forecast from 0.5 the markets show a significantly larger extremeness compared to the average survey ($t(102) = 7.87$, $p < 0.0001$). Overall, market-based forecasts and survey-based forecasts are similar when using a binary metric, however the more extreme market forecasts provide a significantly better predictor when evaluating based on error.

2.4.3 Aggregation Methods

Using alternate aggregations of the individual survey response can create more extreme forecasts which have been linked to better forecasts (29). We provide results for three additional survey aggregation methods: median, simple voting and variance weighted mean. Simple voting includes binarizing every survey response (effectively rounding each response to either 0 or 1) and reporting the percentage of responses which vote for replication success. Variance weighted mean is based on finding a positive relationship between variance in survey responses and overall accuracy. We hypothesize that forecasters with a large variance in survey responses are able to better discriminate between which studies are likely to replicate and which aren't likely, thus providing more extreme forecasts. On the other hand, forecasters who are not able to discriminate provide similar forecasts for many studies, and therefore have a low variance. The median aggregator ($M = 0.63$, $SD = 0.17$), the simple voting aggregator ($M = 0.66$, $SD = 0.21$) and the variance weighted mean ($M = 0.58$, $SD = 0.17$) methods provided higher variance forecasts than the mean ($M = 0.610$, $SD = 0.14$). Evaluating the survey aggregations using mean absolute error simple voting performs the best (0.39), followed variance weighted mean (0.407) and median (0.412). The mean aggregator has the highest mean absolute error of 0.422. The final market price still outperforms these alternate aggregations.

2.4.4 Market Dynamics

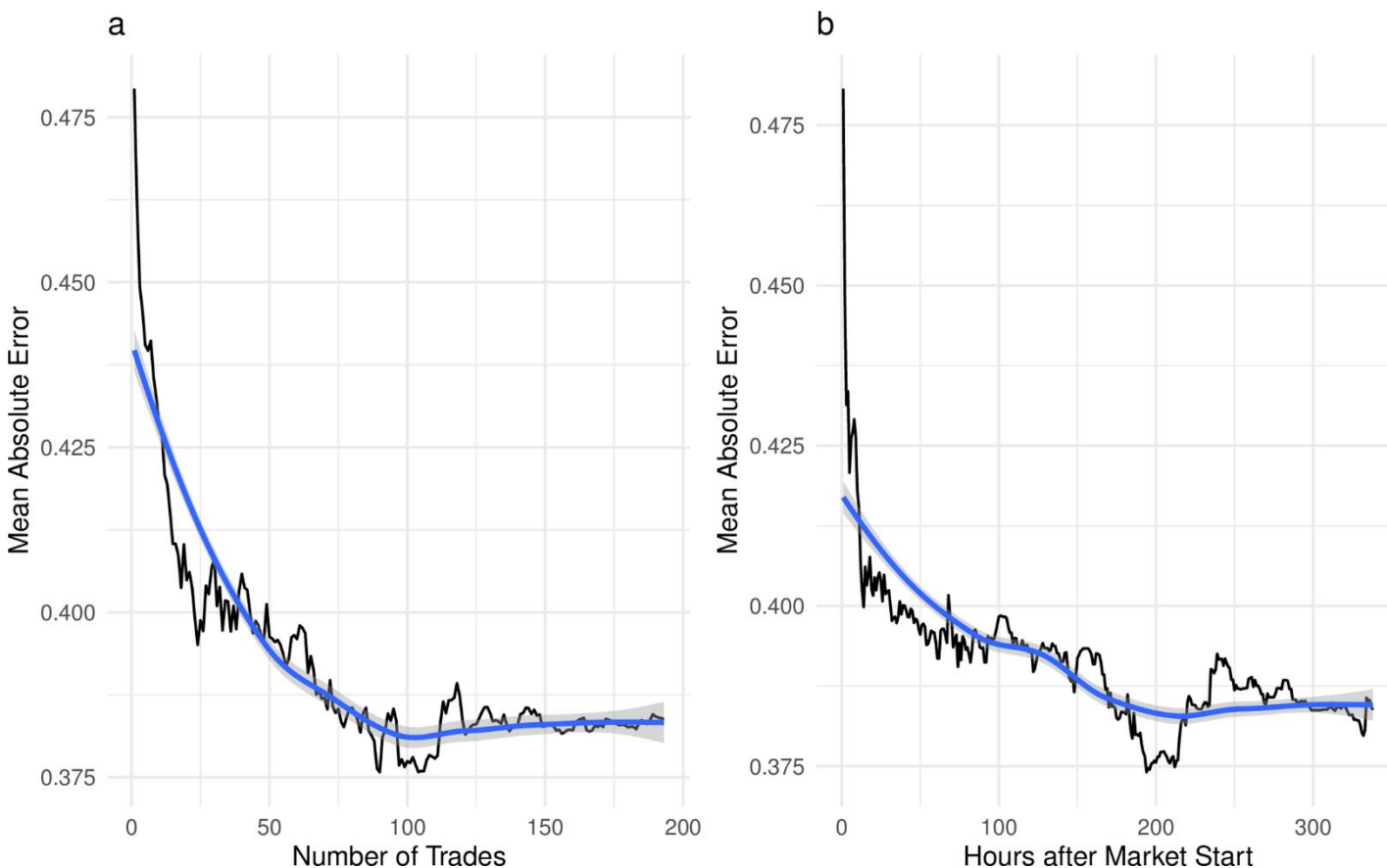
Predictions market are designed to aggregate information that is widely dispersed amongst agents. The market price is expected to converge to a relatively stable value which is interpreted as a probability of the outcome occurring (30,31). For replication markets it is unknown how quickly the market can converge. Using both time and number of trades to quantify how the market progresses, we can investigate on average at what point the information distributed across the agents is 'priced into the market'.

The number of trades in each of the markets range from 26 to 193 ($M = 76$, $SD = 32$) we observe that each forecasting study opened their prediction markets for between 11 and 14 days. The market trades tend to be front loaded, where trading activity diminishes over the available trading time.

As expected, the markets experience diminishing returns in terms of reduction in mean absolute error. Reductions in mean absolute error can be modelled using LOESS regression, where the mean absolute error is estimated at different numbers of trades(32). This model shows that 90% error reduction (i.e 90% of the total error reduction that will occur) happens in the first 69 trades.

When analysing error reduction as a function of time, 65% of the error reduction that will be achieved occurs in the first hour. 90% of total error reduction occurs within the first 161 hours of the markets (just under a week). Both in terms of number of trades and time, the average error fluctuates towards the end the market, without consistently improving forecasts, indicating trades made towards the end of the markets are noisy (Figure 4). However, applying a time weighting smoothing algorithm of all trades after the first week does not result in a significant increase in accuracy.

Figure 2-4. Market Dynamics. Each of the subplots represent reduction in absolute error as the market progresses. Both plots include the mean (across 103 findings) absolute error in black, and the LOESS smoothing in blue. Plot (a) describes error reduction over number of trades and plot (b) describes error reduction over time. We find that the error falls quickly at first, however error does not reduce over

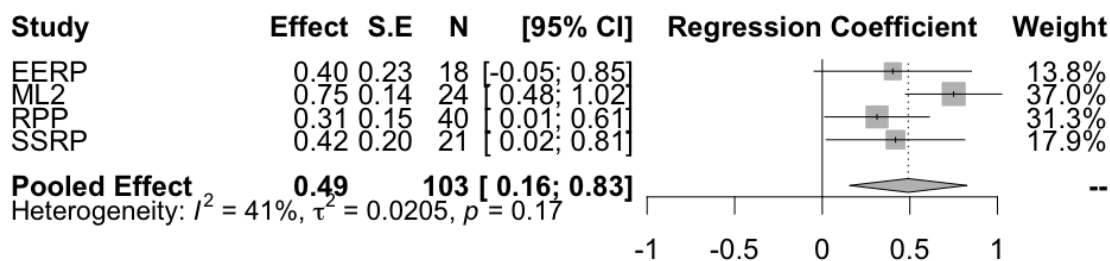


2.4.5 P value Analysis

The p-values of the findings has been shown to be correlated with the replication outcomes (20,21). In particular two other replication based forecasting attempts has shown that p-values are informative to a machine learning algorithm(33,34). We here test this relationship using the pooled market data. One limitation of this analysis is that p-values are often reported as an inequality or a category rather than a real number, for example a typical reported p value is “ $p < 0.05$ ”. Therefore, we transform p-values into categories. As a guide we use categories as suggested by Benjamin et al.

(26); of ‘suggestive evidence’, $p > 0.005$, and statistical significance $p \leq 0.005$. This two-category approach provides a significant relationship ($F(1,101) = 26.515$, $p < .001$, $R^2 = 0.2$) between strength of evidence (through p-values) and replication outcomes ($b = 0.46$, $S.E = 0.089$, $p < .001$). While the replication rate for findings with $p \leq 0.005$ is about 74%, the replication rate for findings with $p > 0.005$ drops to 28%. The correlation of p-value category and outcomes is 0.456 (95% CI 0.29,0.60], $p < .001$). We also study the same effect, by running the same linear model for each study individually, and then combine via a random effects meta-analysis (Figure 5). The meta-analysis also shows a strong relationship between p-value category and replication outcomes (linear model $b = 0.49$, $p=0.019$, 95% CI [0.16,0.83]).

Figure 2-5. P-Value and Replication Outcomes. The relationship between p-value category and replication is robust to a meta-analysis, with a pooled effect of 0.49. There is no evidence for heterogeneity between studies ($I^2 = 41\%$, 95% CI [0%, 80%], $\tau^2 = 0.0205$, 95% CI [0, 0.49] and $Q(3) = 5.06$, $p = 0.17$), however with only 4 studies included in the meta-analysis there may be low power to detect heterogeneity. For all studies, the effect is in the same direction.



2.5 Discussion

In this paper, we investigate the forecasting performance of two different procedures to elicit beliefs about replication of scientific studies: prediction markets and prediction survey. We pooled the forecasting data using these two methods from four published papers in which forecasters, mainly researchers and scholars in the social sciences, estimated the probability that a

tested hypothesis taken from a paper published in scientific journals would replicate. We find that the prediction markets correctly identify replication outcomes 73% of the time (75/103), while the prediction surveys are correct 66% of the time (68/103). Both the prediction market estimates, and the surveys-based estimates are highly correlated with the replication outcomes of the studies selected for replication (Pearson correlation = 0.581 and = 0.564, respectively), suggesting that studies that replicate can be distinguished from studies that do not successfully replicate. However, both the forecasts elicitation methods tend to overestimate the realized replication rates, and beliefs about replication are on average about ten percentage units larger than the observed replication rate. The results suggest that peer beliefs can be elicited to obtain important information about reproducibility, but the systematic overestimation of the replication probability also imply that there is room for calibrating the elicited beliefs to further improve predictions. In terms of comparing which elicitation method performs better in the task of aggregating beliefs and providing more accurate forecasts, our results suggest that the markets perform somewhat better than the survey especially if evaluating based on absolute prediction error.

We confirmed previous results which indicated that p-values, which can be interpreted as a measure for the strength of evidence, are informative in respect to replication success. There is, however, some debate on the appropriateness of interpreting p-values as a measure of strength of evidence (35,36). While Fisher viewed smaller p-values as stronger evidence against the null hypothesis (37), other methods have been proposed to be more suitable for quantifying the strength of evidence (38,39). Our findings thus provide some context for interpreting p-values as strength of evidence by demonstrating a relationship with replicability, but further research could extend this by analysing the relation between replication outcomes with other measures for the strength of evidence such as effect sizes. In addition, a meta-analysis provides no evidence for the relation between the p-value and replication outcomes to differ from project to project (or between

academic fields). Conversely there is suggestive evidence of heterogeneity in the relationship between forecast and replication outcome, as shown by the meta-analysis of the correlations from the different projects. This heterogeneity may arise from differences in study design, the forecasters involved, or some fields may be easier to forecast than others. However, with only a small number of studies used in our meta-analyses, further data are required for more conclusive results.

The data and results presented in this paper can be used for future forecasting projects that are either planned or in progress (14), by informing experimental design and forecasting aggregation. The results can also be used to evaluate the predictive performance of prediction markets against other methods (33,34,40). The pooled dataset presents opportunities for other researchers investigate replicability of scientific research, human forecasts and their intersection, as well as providing a benchmark for any further replication-based markets.

2.6 References

1. Baker M. 1,500 scientists lift the lid on reproducibility. *Nat News*. 2016 May 26;533(7604):452.
2. John LK, Loewenstein G, Prelec D. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychol Sci*. 2012 May 1;23(5):524–32.
3. Ioannidis J. Why Most Published Research Findings Are False. *PLOS Med*. 2005 Aug 30;2(8):e124.
4. Ioannidis J, Doucouliagos C. What's to Know About the Credibility of Empirical Economics? *J Econ Surv*. 2013;27(5):997–1004.
5. Maniadis Z, Tufano F, List JA. One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects. *Am Econ Rev*. 2014 Jan;104(1):277–90.
6. Camerer CF, Dreber A, Forsell E, Ho T-H, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. *Science*. 2016 Mar 25;351(6280):1433–6.
7. Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nat Hum Behav*. 2018 Sep;2(9):637–44.

8. Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, Banks JB, et al. Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J Exp Soc Psychol.* 2016 Nov 1;67:68–82.
9. Klein RA, Ratliff KA, Vianello M, Adams Jr. RB, Bahník Š, Bernstein MJ, et al. Investigating variation in replicability: A “many labs” replication project. *Soc Psychol.* 2014;45(3):142–52.
10. Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB, Alper S, et al. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Adv Methods Pract Psychol Sci.* 2018 Dec;1(4):443–90.
11. Landy J, Jia M, Ding I, Viganola D, Tierney W, Dreber A, et al. Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychol Bull* 2019 Oct 29 [cited 2020 Jan 20]; Available from: <http://repository.essex.ac.uk/25784/>
12. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science.* 2015 Aug 28;349(6251):aac4716.
13. Schweinsberg M, Madan N, Vianello M, Sommer SA, Jordan J, Tierney W, et al. The pipeline project: Pre-publication independent replications of a single laboratory’s research pipeline. *J Exp Soc Psychol.* 2016 Sep 1;66:55–67.
14. Gordon M, Viganola D, Bishop M, Chen Y, Dreber A, Goldfedder B, et al. Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *R Soc Open Sci.* 2020 Jul 22;7(7):200566.
15. Christensen G, Miguel E. Transparency, Reproducibility, and the Credibility of Economics Research. *J Econ Lit.* 2018 Sep;56(3):920–80.
16. Etz A, Vandekerckhove J. A Bayesian Perspective on the Reproducibility Project: Psychology. *PloS One.* 2016;11(2):e0149794.
17. Fanelli D. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proc Natl Acad Sci.* 2018 Mar 13;115(11):2628–31.
18. Pashler H, Harris CR. Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspect Psychol Sci.* 2012 Nov 1;7(6):531–6.
19. Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, et al. Using prediction markets to estimate the reproducibility of scientific research. *Proc Natl Acad Sci.* 2015 Dec 15;112(50):15343–7.
20. Forsell E, Viganola D, Pfeiffer T, Almenberg J, Wilson B, Chen Y, et al. Predicting replication outcomes in the Many Labs 2 study. *J Econ Psychol.* 2019 Dec 1;75:102117.
21. Cumming G. Replication and p Intervals: p Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspect Psychol Sci.* 2008 Jul 1;3(4):286–300.
22. Ioannidis JPA. Why Most Discovered True Associations Are Inflated: *Epidemiology.* 2008 Sep;19(5):640–8.

23. Manski CF. Interpreting the predictions of prediction markets. *Econ Lett.* 2006 Jun 1;91(3):425–9.
24. Hanson R. Combinatorial Information Market Design. *Inf Syst Front.* 2003 Jan 1;5(1):107–19.
25. Forsell E, Viganola D, Pfeiffer T, Almenberg J, Wilson B, Chen Y, et al. Predicting replication outcomes in the Many Labs 2 study. *J Econ Psychol.* 2018 Oct 25 [cited 2019 Mar 22]; Available from: <http://www.sciencedirect.com/science/article/pii/S0167487018303283>
26. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. *Nat Hum Behav.* 2018 Jan;2(1):6.
27. Harrer M, Cuijpers P, Furukawa T, Ebert D. Doing meta-analysis in R: A hands-on guide. *Prot Lab Erlangen.* 2019; Available from: https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/
28. Int'Hout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol.* 2014 Feb 18;14(1):25.
29. Baron J, Mellers BA, Tetlock PE, Stone E, Ungar LH. Two Reasons to Make Aggregated Probability Forecasts More Extreme. *Decis Anal.* 2014 Mar 19 [cited 2019 Mar 20]; Available from: <https://pubsonline.informs.org/doi/abs/10.1287/deca.2014.0293>
30. Arrow KJ, Forsythe R, Gorham M, Hahn R, Hanson R, Ledyard JO, et al. The Promise of Prediction Markets. *Science.* 2008 May 16;320(5878):877–8.
31. Atanasov P, Rescober P, Stone E, Swift SA, Servan-Schreiber E, Tetlock P, et al. Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls. *Manag Sci.* 2017 Mar;63(3):691–706.
32. Cleveland WS, Devlin SJ. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *J Am Stat Assoc.* 1988 Sep 1;83(403):596–610.
33. Yang Y, Youyou W, Uzzi B. Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proc Natl Acad Sci.* 2020 May 19;117(20):10762–8.
34. Altmejd A, Dreber A, Forsell E, Huber J, Imai T, Johannesson M, et al. Predicting the replicability of social science lab experiments. *PLOS ONE.* 2019 May 12;14(12):e0225826.
35. Wicherts JM, Bakker M, Molenaar D. Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. *PLOS ONE.* 2011 Feb 11;6(11):e26828.
36. Gibson EW. The Role of p-Values in Judging the Strength of Evidence and Realistic Replication Expectations. *Stat Biopharm Res.* 2020 Jun 23;0(0):1–13.
37. Fisher R. *Statistical Methods for Research Workers.* Hafner PublishingCo. Inc N Y. 1958;212–47.

38. McBride G, Cole RG, Westbrooke I, Jowett I. Assessing environmentally significant effects: a better strength-of-evidence than a single P value? *Environ Monit Assess.* 2014 May 1;186(5):2729–40.
39. Goodman SN. Introduction to Bayesian methods I: measuring the strength of evidence. *Clin Trials.* 2005 Aug 1;2(4):282–90.
40. Pawel S, Held L. Probabilistic forecasting of replication studies. *PLOS ONE.* 2020 Apr 22;15(4):e0231416.

3 Chapter 3: SCORE Pre-registration

There are three pre-registration documents included in this chapter; pre-registration of the analysis of meta markets and surveys (the results of which are found in Chapter 4), pre-registration of the analysis of regular markets and surveys (overview found in Appendix 1) and finally the pre-registration of the analysis of a COVID-19 related round. Each pre-registration includes experimental design and statistical models as well as methodology for calculating Confidence Scores (CS). Confidence Scores are the key output of our research and indicate the probability a scientific claim will be successfully replicated. Some of the confidence scores rely on a method of peer assessment called ‘Surrogate Scoring Rules’ or SSR. The design and implementation of SSR was completed by my colleagues, and therefore I was not involved in the calculation of confidence scores relating to SSR. I did however, calculate all other confidence scores. The final round of forecasting was focused on social and behavioural claims related to COVID-19 and was not in the initial plan for SCORE, but was a later addition in response to the pandemic. These 100 claims are not included in the 3000 claims assessed as part of the regular SCORE programme and also differ from the sample of 400 COVID related preprints as described in Chapter 7. I contributed to the experimental and statistical design and the drafting of this preprint. In addition I also developed the methodology for many of the confidence scores.

Due to the COVID-19 pandemic, the replications of scientific claims by TA1 team (see Appendix 1) were delayed, resulting in no validation available to perform the hypotheses found in these pre-registration documents. Therefore, these pre-registration documents represent research that was completed by our Replication Markets Team (also referred to as “Team KeyW” in the document below).

To align with the formatting and referencing style of this thesis, there are some changes in formatting and referencing style of the original document

Pre-registration documentation for SCORE TA2/Team KeyW

3.1 Context

With its SCORE programme, the Defense Sciences Office (DSO) at the Defense Advanced Research Projects Agency (DARPA) is funding innovative research projects for the development and deployment of tools to assign Confidence Scores (CSs) to different kinds of Social and Behavioral Science (SBS) research results and claims. CSs are quantitative measures that should enable someone to understand the degree to which a particular claim or result is likely to be reproducible and/or replicable. As part of the SCORE project, our team (team KeyW) will develop surrogate scoring and prediction market approaches to elicit CSs from market participants. A total of 3,000 claims will be scored. A database with claims and descriptive data (CVD) will be created by a third party (TA1; see “Data Integration Plan”, DARPA SCORE Internal Documentation, 2019). TA1 will also coordinate the validation of claims through direct replication, and through data-analytic reproduction using existing alternative data. The number of claims validated through direct replication will be around 100 (3.3%) and an additional 100 claims will be validated through data-analytic reproduction. Our team will not receive information about which claims will be selected for reproduction or replication, and will not receive claim-specific detail on replications. The outcome of the TA1 replications will be used to evaluate the scores provided by our team. This evaluation is conducted by an external party (T&E; see “Phase 1 TA1 and TA2 Test and Evaluation (T&E) Plan”, (1)). Primary performance metrics will be (a slightly modified version of) Wilcoxon-Mann-Whitney’s U. 10 CS scores will be submitted.

3.2 Methodology

3.2.1 Overview

Team KeyW is focusing on two methodologies to elicit forecasts on the replicability of claims: decision markets and surrogate scoring. Decision markets are mechanisms related to prediction markets that are modified to provide incentivized forecasts despite the low rate of resolution. Surrogate scoring is a survey-based incentivized elicitation method that does not require access to a ‘ground truth’ to determine incentives. Of the 10 scores, 3 will be based on the decision markets, 6 will be based on the surrogate scoring surveys, and one will use information from both approaches. A description of these scores is given in Section 5.

3.2.2 Forecasting in monthly rounds

Forecasts on the replicability of the 3,000 claims are elicited in monthly rounds, with each round covering about 300 claims. Each round will start with surveys for the surrogate scoring method. Each participant will be assigned to a survey with questions on 10 of the 300 claims in that round. The assignment is done based on preferences as declared in an entry survey. Once participants complete this first batch they can choose to complete a different ‘batch’ of claims. Claims within a batch are, if feasible, sampled from the same journal. The surveys will be accessible for one week. Once the survey closes, markets on the replicability of the 300 claims will open for two weeks. Subsequent rounds will be conducted in a 4-week cycle. Participants are recruited through a variety of community mailing lists and social media; survey and markets can be accessed through the website www.replicationmarkets.com. Once registered, they are asked to complete compulsory and non-compulsory intake surveys. Before starting with the regular monthly forecasting rounds, there will be a survey and market on meta-questions.

3.2.3 Meta-questions

Before eliciting forecasts for individual claims, a market and corresponding survey for meta-questions (referred to as meta-survey and meta-market) will be used to forecast overall replication rates in SCORE, subject-specific replication rates and year-wise replication rates. The questions will be the same for survey and market and are given in [Appendix A](#). These meta-questions allow to test whether participants expect field-specific, and time-dependent variation of the replication rates, and will help calibrating claim-level forecasts in the regular rounds. The meta-survey opens on August 12, 2019, noon UTC, and closes August 18, 2019, 11:59pm UTC. A surrogate method is used to incentivize participants and aggregate survey responses of those participants that are expected to be most accurate. The methodology will be the same as for the score *CS_survey_Q1* and *CS_survey_Q1_A1* (see Section 5 for details). The meta-market will open for the subsequent week (August 19, noon – August 25, 11:59pm UTC), and might be re-open periodically throughout the project period. The market platform uses a market maker implementing a logarithmic market scoring rule (LMSR) with a base of 2, and a liquidity parameter of $b = 100$. All meta-markets start with an initial pricing of 0.5, and will settle at the respective observed replication rate. Participants will receive 100 points to invest; points that have not been invested by the closing time of the meta-market will be lost. For the pre-registration of the statistical analysis of the data from the meta-questions see Section 3.

3.2.4 Surrogate scoring⁵

In the regular monthly rounds, surrogate scoring will be used to elicit forecasts on the outcome of direct replications for the individual claims. In addition, the survey questions will address belief on the beliefs of other forecasters, the outcome of data-analytic reproduction, the plausibility of claims, the likelihood of claims to be evaluated through direct replication, and errors in the claim data (see [Appendix B](#)). Survey responses will be elicited for batches of 10 claims (see above). Incentives will be provided only for forecasts on the outcome of direct replications (SQ1) and data-analytic reproductions (SQ3). The surrogate scoring methodology will be used to identify and incentivize the expected top forecasters in each round for each batch. The surveys will be open for one week at the beginning of each round. The aggregation of survey responses into forecasts is described in Section 5.

3.2.5 Prediction markets

Once the surrogate scoring survey is closed, the market for the replicability of the 300 individual claims in the round will open. The main market-based mechanism to elicit forecasts for the outcome of the individual (claim-level) replications will be through the trading of ‘shares’ on binary (yes/no) questions. For each claim in the CVD database, there will be ‘Yes’ shares and ‘No’ shares. ‘Yes’ shares will pay 1 point if the underlying claim is evaluated through direct replication, and the replication yields a statistically significant finding in the direction of the original claim. ‘No’ shares will pay 1 point if underlying claim is evaluated through direct replication, and the replication does not yield a statistically significant finding in the direction of the original claim. If a claim is not

⁵ Surrogate Scoring for prizes and confidence scores were calculated by other members of the replication markets (or Team KeyW). I was not involved in this process.

selected for replication, both ‘Yes’ shares and ‘No’ shares will pay out 0 points. This approach is equivalent to using decision markets with simple stochastic decision rules. Trading will be facilitated through a market maker implementing a logarithmic market scoring rule (LMSR) with base 2 and a liquidity parameter of $b = 100$. Initial decision market prices will be informed by p-values and the relation between p-value and replication rates as observed in previous replication markets. Based on the field ‘coded p-value’ in CVD, each claim will be assigned into one of the following three categories (“ $p \leq 0.001$ ”, “ $0.001 < p \leq 0.01$ ”, “ $p > 0.01$ ”). Starting prices for these three categories will be (0.8, 0.4, 0.3)⁶. Participants receive 300 points, and can obtain a bonus of up to 30% depending on ongoing contributions to the market and survey. In each round the market will be open for 2 weeks. Points that are not invested by the time the market closes will be lost. Forecasts will be generated by smoothing the market prices over the second half of the trading period (see Section 5, *CS_market_smoothed* for details).

3.2.6 Prizes

The total prize pool will be split into one part dedicated to the prediction markets (~two thirds), and one part dedicated to surrogate scoring (~one third). Replication results are expected to be released at the end of the project. The prize pool for the prediction markets will thus be paid out at the end of the project. Prizes for surrogate scoring will be paid out after each round. The payouts for surrogate scoring will be made after the markets on a round close. This ensures that information on performance in surrogate scoring does not affect investments in the market. The prize pool for surrogate scoring will be allocated into prize pools of USD 160 for the individual batches. The top four participants for any given batch (as determined by their average score over the round; see Section

⁶ The rationale for these starting prices can be found in chapter 2

5 for details) will receive fixed prizes of USD 80, USD 40, USD 20, and USD 20. The prize pool for the monthly prediction markets will be allocated into separate pools of about USD 9,000 for each round (USD 900 per resolving claim, with 10 resolutions expected per round). This prize pool will be allocated proportionally to the points that participants earn from their investments. Participants in the prediction markets do not receive any pay outs for points that have not been invested into forecasts.

The prize pool for meta-surveys is USD 960, and is allocated in prizes of USD 80 for the top 6 participants, USD 40 for participants ranked 7-12, and USD 20 for participants ranked 13-24. The prize pool for the meta-market is USD 4,320 (USD 360 for each of the 12 meta-claims) and is allocated proportionally to the points earned in these markets.

3.3 Pre-registration for the statistical analysis of the meta-markets and meta-surveys

In the meta-market and meta-surveys we use prediction markets and surveys to elicit ‘meta-forecasts’ on the following replication rates:

- the overall replication rate in the SCORE project;
- the topic-specific replication rates;
- the year-specific replication rates.

The questions are given in [Appendix A](#). In this section, we specify the analyses we plan to perform. Unless otherwise specified, we interpret the threshold of $p < 0.005$ as identifying statistical significance and the threshold of $p < 0.05$ as identifying suggestive evidence; all the tests in this pre-analysis plan are two-sided tests. The hypotheses are expressed in directional terms unless the direction of the tested effect is not clear ex-ante, in which case they are expressed in non-directional

terms. For the market transaction data, no data will be excluded. For the meta-survey data, the responses of participants who skipped more than 6 of 12 questions will be excluded.

3.3.1 Primary Hypothesis

1. Do forecasters in their meta-survey response on the overall replication rate (MSQ1) in SCORE over- or underestimate overall observed replication rates?

Test: Independent samples t-test between a vector containing the participants' survey responses for the overall replication rate, and a vector containing the outcomes of the (approximately) 100 direct replications, with 1 denoting a replication outcome consistent with the original effect (for definition see "Phase 1 TA1 and TA2 Test and Evaluation (T&E) Plan", DARPA SCORE Internal Documentation, 2019), and 0 denoting a replication outcome not consistent with the original effect. The sample variances are treated as being different, i.e. an independent samples t-test not assuming equal variances will be used.

2. Is there evidence for topic-specific replication rates to differ between topics?

Test: We use a one-way repeated measures ANOVA across the participants' survey responses (MSQ2-MSQ7) as dependent variable, and topic (i.e. question ID) as independent factor. If the model indicates that there is a joint significant effect of the topic variables, we compute pairwise paired t tests between responses to each topic, and apply the Benjamini-Hochberg procedure to control the false discovery rate.

3. Is there evidence for survey responses on year-specific replication rates to differ between the time periods?

Test: We use a one-way repeated measures ANOVA across the participants' survey responses (MSQ8-MSQ12) as dependent variable, and year-band (i.e. question ID) as independent factor. If the model indicates that there is a joint significant effect of the year-band variables, we compute pairwise paired t tests between responses to each year-band, and apply the Benjamini-Hochberg procedure to control the false discovery rate.

4. Do topic-specific forecasts depend on the field of the forecaster?

Test: Based on survey responses to (MSQ2-MSQ7) we use a paired t-test between the participants' response for the expected replication rate in their own field, and the average of their responses for the expected replication rate in the other fields. To determine participants' fields, we use responses to the question "Which field describes best your work and/or interest" in the demographic survey. For participants who select multiple options on this question, we use the average of the expected replication rates for these fields as the value for their own field. Participants that do not respond to this question will be excluded from this analysis.

5. Does the error of topic-specific forecasts depend on the field of the forecaster?

Test: Based on survey responses to (MSQ2-MSQ7) we use a paired t-test between the participants' error for the expected replication rate in their own field, and the average of their error for the expected replication rate in the other fields. To determine participants' fields, we use responses to the question "Which field describes best your work and/or interest" in the demographic survey. For participants who select multiple options on this question, we use the average of the errors for these fields as the value for their own field. Participants that do not respond to this question will be excluded from this analysis. To determine the error, we use the absolute difference of a participant's response (MSQ2-MSQ7) and the observed replication rate for the claims that fall into this topic.

3.3.2 Secondary Hypotheses

6. Survey-based aggregated forecasts and market-based forecasts are correlated

Test: The survey-based aggregated forecasts on topics and time periods are merged into a single vector. Similarly, the smoothed market-based forecasts on topic and time periods are merged into a single vector. We test for a correlation of both vectors, using Pearson's correlation coefficient.

7. Observed outcomes are correlated with aggregated survey-based forecasted outcomes

Test: The survey-based aggregated forecasts on topics and time periods are merged into a single vector. A second vector contains the observed replication rates for each time period and topic. We test for a correlation of both vectors, using Pearson's correlation coefficient.

8. Observed outcomes are correlated with smoothed market-based forecasts

Test: The market-based aggregated forecasts on topics and time periods are merged into a single vector. A second vector contains the observed replication rates for each time period and topic. We test for a correlation of both vectors, using Pearson's correlation coefficient.

3.4 Pre-registration for the statistical analysis of the regular markets and surveys

Unless stated otherwise, we only use survey responses of participants who responded to at least 5 claim-level surveys. Unless otherwise specified, we interpret the threshold of $p < 0.005$ as identifying statistical significance and the threshold of $p < 0.05$ as identifying suggestive evidence; all the tests in this pre-analysis plan are two-sided tests.

3.4.1 Performance of the scores

3.4.1.1 Primary Hypotheses

1. Do the CSs (see Section 5; all scores are used) overestimate or underestimate actual replication rates?

Tests: Paired t-test between binary outcome data (1 = replicated, 0 = not replicated) and each of the computed CSs. Only the outcomes of the **replications** and corresponding forecasts will be used.

Priority for reporting will be given to survey means (CS_survey_Q1_mean), the surrogate score CS_survey_Q1, and the raw market forecast (CS_market_raw). The results for the other scores will be treated as secondary/exploratory hypotheses.

2. How do the scores differ from each other in terms of forecasting performance. Which CS(s) perform best?

2. a. Evaluation based on **replication** results.

Tests: Brier scores are calculated for each of the CSs with observed **replication outcome**. We then use pairwise paired t-tests of the Brier scores to test for differences in the mean. To control the false discovery rate, we use the Benjamini – Hochberg correction.

We expect that the final market prices are improved over the initial market prices, and that smoothing and calibration further improves forecasts. For the survey-based forecasts, we expect that all surrogate scoring methods are improved over simple survey averages. It is plausible that merging market forecasts with survey-based forecasts as outlined in Section 5 leads to further improvement over both the calibrated market scores and the calibrated surrogate scoring surveys. Priority for reporting will thus be given to comparisons between **market based scores** (initial score *CS_initial* vs. raw market price *CS_market_raw*; raw market price *CS_market_raw* vs smoothed market price *CS_market_smoothed*; smoothed market *CS_market_smoothed* vs calibrated market

CS_market_calibrated; and calibrated market *CS_market_calibrated* vs merged survey and market score *CS_survey_market*), and between **survey based scores** (survey mean *CS_survey_Q1_mean* vs surrogate scores *CS_survey_Q1*, surrogate scores *CS_survey_Q1* vs calibrated surrogate scores *CS_survey_Q1_calibrated*, and calibrated surrogate scores *CS_survey_Q1_calibrated* vs merged survey and market score *CS_survey_market*). All other comparisons will be treated as secondary/exploratory hypotheses.

3.4.1.2 Secondary/Exploratory Hypotheses

2. b. Evaluation based on **reproduction** results.

Tests: Brier scores are calculated for each observed **data-analytic reproduction outcome** and each of the CSs. We then use pairwise paired t-tests of the Brier scores to test for differences in the mean. To control the false discovery rate, we use the Benjamini – Hochberg correction.

2. c. Evaluation based on **combined replication and reproduction** results.

Tests: Brier scores are calculated for **combined replication and reproduction outcome** and each of the CSs. We then use pairwise paired t-tests of the Brier scores to test for differences in the mean. To control the false discovery rate, we use the Benjamini – Hochberg correction.

3.4.1.3 Secondary/Confirmatory Hypothesis

3. **Correlation between the 10 CSs which are reported to DARPA (“yes” in *Included in 10 CSs* column in Section 5)**

Tests: We calculate Pearson correlation coefficients between for each pair of CSs. The scores for all (approximately 3000) claims are included for these tests.

3.4.1.4 Additional descriptive analyses

We calculate for all scores (including those not reported to DARPA) the mean absolute forecasting error, mean Brier score (including decompositions), AUC, and the R^2 , intercept and coefficient of a linear model between CSs and the observed replication outcomes. This work is in addition to the evaluation undertaken by T&E (see “Phase 1 TA1 and TA2 Test and Evaluation (T&E) Plan”, DARPA SCORE Internal Documentation, 2019).

3.4.2 Participant characteristics

3.4.2.1 Primary Hypotheses

4. Which demographic variables are associated with forecasting accuracy?

Tests: For each participant, and each claim validated through replication, we calculate the Brier score for the response to survey Question Q1. We then conduct individual linear regressions of the Brier score and each of the demographic variables listed below. This is followed by single linear regression model for the Brier score and all demographic variables. For all models, clustering the standard errors of participants is used.

Demographic variables:

- 1) Career stage in University (Undergrad, Grad student, etc., as dummy variables)
- 2) Index created as the average of self-reported expertise in mathematics, quantitative modeling, statistics, probability, experimental design, risk analysis, forecasting
- 3) Actively open-minded index
- 4) Berlin numeracy test score
- 5) Index based on the % of correct answers about the p-values
- 6) Index based on the previous replication quiz (% of correct guesses)

5. Are participants more accurate in their own field as opposed to other fields?

Tests: Paired t-test between participants’ mean Brier score of survey forecasts where the field of claim matches fields of interests, and participants’ mean Brier score of survey forecasts where

field of claim does not match fields of interests. We only include participants who provide at least 5 survey forecasts to claims that match fields of interest AND 5 survey forecasts to claims that do not match fields of interest.

3.4.2.2 Secondary Hypotheses

6. Which demographic variables are correlated with market returns?

Tests: For each participant, we calculate the average market return per round, and use this as dependent variable in a linear regression model. As independent variables, we use the demographic variables listed for Hypothesis 4. As for Hypothesis 4, we use individual linear regressions for each of the demographic variables, and one regression model that includes all demographic variables.

7. Does Surrogate Scoring identify the best forecasters?

Tests: Pearson's correlation test between the participants' mean surrogate score and the mean Brier scores calculated based on observed replication outcomes.

8. Do more accurate forecasts come from more diverse batches?

Tests: The Brier score of surrogate score *CS_survey_Q1* calculated based on observed replication outcomes is correlated with the Gini-Simpson index of the diversity, in terms of fields, of the forecasters who contributed to the forecast. We use a Spearman correlation for this test. The test is repeated with the mean survey response *CS_survey_Q1_mean* instead of *CS_survey_Q1*.

9. Is the average time a participant spends per claim to answer the surveys correlated with the participant's mean Brier score?

Tests: Spearman correlation between median time per claim spent completing surveys and the participant's average Brier score calculated using observed replication outcomes and the response given to survey question Q1.

10. Is the number of claims a participant provides forecasts for in the surveys correlated with their mean Brier Score?

Tests: Spearman correlation between the number of claims a participant respond to, and the participant's mean Brier scores, calculated using observed replication outcomes and the response given to survey question Q1.

11. Is the variation of a participant's survey responses correlated with their overall forecasting accuracy?

Tests: For each participant, we calculate the variation over all responses given to Q1, and the mean Brier score using response to Q 1 and observed replication outcomes. We use Pearson's correlation coefficient to test for a correlation between these two variables.

3.4.3 Study characteristics

3.4.3.1 Primary Hypotheses

12. Is there a relation between the p-value of the original study and the rate of replication?

Tests: We use a linear model with replication outcome as dependent variable, and original study p-value category as independent variable. We use the same categorization as used for the initial market prices.

13. Is p-value impact on correlation similar to pooled past replication markets?

Tests: Fisher's exact test between numbers of successful and unsuccessful replications in each p-value category found in previous studies (pooled market data from past project RPP, SSRP, EERP, and ML2; manuscript in preparation) vs numbers of successful and un-successful replications in the SCORE claim database in each p-value category. In total, three tests are conducted, one for each p-value category. We exclude claims for which no p-value could be identified in the CVD database.

3.4.3.2 Secondary Hypotheses

The following tests (hypothesis 14-15) are conditional on whether standardized figures are delivered by TA1.

14. Is original claim effect size (as stated in CVD database) correlated with replication effect size?

Tests: Spearman correlation between original claim effect size and replication outcomes.

15. Is original claim sample size (as stated in the CVD database) correlated with replication outcome?

Tests: Spearman correlation between original claim sample size and replication outcome.

16. Is original claim effect size (as stated in the CVD database) correlated with final market price?

Tests: Spearman correlation between original claim effect size and final market price (*CS_market_raw*).

17. Is the original claim sample size (as stated in the CVD database) correlated with final market price?

Tests: Spearman correlation between original claim sample size and final market price (*CS_market_raw*)

18. Are effect size and sample size correctly priced into the market

Tests: We will conduct the following logistic regression models:

1. Final Market Price against outcome
2. Final Market Price + standardized effect size against outcome
3. Final Market Price + standardized effect size + standardized sample size against outcome

We then apply the likelihood ratio test between the above models to understand if the difference in model fit between the complex and simple models is statistically significant.

3.4.4 Comparison with Round 0, market dynamics, and other areas

3.4.4.1 Secondary Hypotheses

19. Are aggregated claims-level forecasts in the market/survey are consistent with meta markets/surveys?

Tests: For each forecast from the meta-market (i.e. for 5 year-bands, 6 topics, and the overall replication rate in score), we use one-sample t-tests to test for a difference between meta-market forecast and the mean of the corresponding claim-level market forecasts *CS_market_smoothed*. These tests are repeated for the meta-survey forecasts, and the corresponding claim-level survey forecasts *CS_survey_Q1*.

This test will help to determine the effectiveness of calibration.

20. Are number of trades correlated with the Brier score of final market price?

Tests: We use Pearson's correlation coefficient to test for a correlation between the number of trades in the market, and the Brier score for forecasts from *CS_market_raw* calculated based on the replication outcomes.

21. Is traded volume (in terms of shares traded) correlated with the Brier score for the market forecast

Tests: We use Spearman's correlation coefficient to test for a correlation between the number of traded shares for a claim, and the Brier score for forecasts from *CS_market_raw* calculated based on the replication outcomes.

22. Do participants correctly anticipate which claims will be replicated or reproduced?

Tests: Paired t-test between response to survey Question 6 forecasted and a binary value if a claim was selected for replication (1 = selected for replication, 0 = not selected for replication). This test is repeated for reproduction, where the binary variable reflects selection for reproduction rather than replication.

23. Does final volume of holdings differ for validated claims vs. non-validated claims?

Tests: Unpaired t-tests of volume holdings (measured in traded shares) for validated claims vs unvalidated claims.

This test is suited to detect if participants can detect which claims are more likely validated.

3.4.4.2 Descriptive analyses

We use descriptive analyses to study at what time does the market reach 95% in error reduction; after how many trades does the market reach 95% in error reduction; how many trades there are per participant; and to study the distribution of final positions.

Disclaimer: This is a long-term project resulting in a complex dataset. We reserve the right to change experimental approaches to adjust to experimental opportunities and constraints. Changes will be documented in amendments to this document. We also reserve the right to adjust analyses and to test additional hypotheses. This will be indicated in the resulting documentation.

3.5 Scores

The table below summarizes the scores calculated by Team KeyW, and indicates which ones are reported to DARPA. Key variables include:

- From **Meta-markets and meta-surveys**, the variables $meta_market_y$ and $meta_market_t$ denote the replication rate of a claim as expected for the time period of publication, and for the topic. Smoothing is applied using the same method as described for $CS_market_smoothed$. The variables $meta_survey_y$ and $meta_survey_t$ denote the corresponding estimates from the meta-surveys, using surrogate scoring aggregation as described for CS_survey_Q1 .
- From the **regular (claim-level) markets**, variables $t_j(n)$ and $PR_j(n)$, denote the time of trade number n in claim j , and the updated price after this trade.
- From the **claim-level surveys**, $SQ1_{i,j}$, $SQ2_{i,j}$, ..., denote the response of participant i to Question Q1, Q2, ..., in the surrogate scoring survey for claim j .

Table 3-1: Confidence Score Descriptors

Name	Included in the 10 CSs	Short description

<i>Market-based scores</i>		
<i>CS_market_raw</i>	Yes	<p>Closing price of the markets on the binary claims. The market platform uses a logarithmic market scoring rule (LMSR) to update prices as shares are traded. After trade n in the shares of a CVD claim j, the market price for this share is updated from price $PR_j(n-1)$ to $PR_j(n)$. The score $CS_market_raw_j$ is the updated price after the final trade on claim j. This is analogous to what has been used in previous replication markets.</p>
<i>CS_market_smoothed</i>	Yes	<p>Smoothed market score. In previous replication markets 95% in the reduction of the forecasting errors occurs in the first 31 hours (56 trades) of the market. Afterwards there is little improvement in accuracy, and trading appears to be more noisy than informative. We aim to reduce the effect of noise by averaging the prices PR_j for a claim j over the second half of the trading period. To do so, we find the last trade k in the first half of the trading period and the last trade m in the second half of the trading period, and calculate</p> $CS_market_smoothed_j = \sum_{n=k..m} \frac{PR_j(n)(t_j(n+1) - t_j(n))}{t_j(m+1) - t_j(k)}$ <p>Here, $t_j(n)$ denotes the time of trade n on claim j; $t_j(m+1)$ denotes the closing time of the market. The time period used for averaging will be a subject of round-by-round review and might be adjusted depending on the distribution of trades over time.</p>
<i>CS_market_calibrated</i>	Yes	<p>Smoothed and calibrated market score. Previous replication markets have shown that a calibration of market prices can improve forecasts: p-values are not sufficiently priced in, and overall markets tend to be too optimistic. Forecasts from smoothed market prices can thus likely be improved by using information from the meta-markets, p-value, and past projects. We adjust the smoothed market score for each claim based on the CVD p-value, a topic-specific correction, and a time-specific correction. The correction will be applied in log-odds space to ensure the resulting scores remain between zero and one.</p> $\begin{aligned} & \text{logit}(CS_market_calibrated_j) \\ & = \text{logit}(CS_market_smoothed_j) + \Delta^P_j + \Delta^T_j + \Delta^Y_j \end{aligned}$ <p>The correction terms are calculated as</p>

		$\Delta_j^P = \text{logit}(CS_initial_j) - \text{mean_logit_pval}_j$ $\Delta_j^Y = \text{logit}(meta_market_y_j) - \text{mean_logit_y}_j$ $\Delta_j^T = \text{logit}(meta_market_t_j) - \text{mean_logit_t}_j$ <p>Here, $meta_market_y_j$ and $meta_market_t_j$ denote the estimates from the meta-markets for replication rates from publications of the same time period, and same topic, as claim j, respectively. Additionally, $mean_logit_pval_j$, $mean_logit_y_j$, and $mean_logit_t_j$, are the averages of the logit-transformed market estimates for the replication probabilities for claims that fall into the same category in terms of p-value, time period, and topic, as claim j, respectively.</p>
Survey-based scores		
<i>CS_survey_Q1_mean</i>	No	Average response to survey question Q1 over all participants
<i>CS_survey_Q1</i>	Yes	<p>Forecast from survey responses as elicited in survey question Q1, using the aggregation method with best track-record on similar data, the surrogate-score-aided aggregator. A surrogate score, using the Brier Score as the underlying metric, is computed and maintained for each survey participant to track their performance.</p> <p>To compute such score for each response of Q1, we first construct a (random) reference answer Y_j' for each claim j. Y_j' is a binary random variable following a Bernoulli distribution with parameter $\text{Mean}_{\text{over}_j}(SQ1_{i,j})$.</p> <p>Our estimation algorithms will then compute two hyper-parameters w_1, w_0. Then for each response $SQ1_{i,j}$, the surrogate score under our constructed reference Y_j' is</p> $S'(SQ1_{i,j}, Y_j') = w_{Y_j'} * \text{Brier}(SQ1_{i,j}, Y_j') + w_{1-Y_j'} * \text{Brier}(SQ1_{i,j}, 1-Y_j').$ <p>We compute the expected surrogate score for $SQ1_{i,j}$ as</p> $ES_{i,j} = \mathbb{E}_{Y_j' \sim \text{Bern}(\text{Mean}_{\text{over}_i}(SQ1_{i,j}))} [S'(SQ1_{i,j}, Y_j')]$ <p>For each participant i, we then take the mean of $ES_{i,j}$:</p> $MES_i = \text{Mean}_{\text{over}_j}(ES_{i,j})$ <p>On each claim, we select answers from the top ($\max(5, 10\% * \text{number of forecasts on the forecast question})$) participants w.r.t. MES_i up to date and perform mean aggregation over their forecasts on that claim. For further information on the methodology see Liu, Y., & Chen, Y. (2018). Surrogate Scoring</p>

		Rules and a Dominant Truth Serum. <i>arXiv preprint arXiv:1802.09158</i> .
<i>CS_survey_Q1_calibrated</i>	Yes	<p>Calibrated CS_survey_Q1 scores. Forecasts from <i>CS_survey_Q1</i> may also likely be improved by using information from the meta-surveys, p-values, and past projects. We have:</p> $\text{logit}(CS_survey_Q1_calibrated_j)$ $= \text{logit}(CS_survey_Q1_AI_j) + \Delta^P_j + \Delta^Y_j + \Delta^T_j$ $\Delta^P_j = \text{logit}(CS_initial_j) - \text{mean_logit_SSR_pval}_j$ $\Delta^Y_j = \text{logit}(meta_survey_y_j) - \text{mean_logit_SSR_y}_j$ $\Delta^T_j = \text{logit}(meta_survey_t_j) - \text{mean_logit_SSR_t}_j$ <p>Here, <i>meta_survey_y_j</i>, <i>meta_survey_t_j</i> denote the estimates from the meta-surveys for replication rates from publications of the same time period, and same topic, as claim <i>j</i>, respectively. Meanwhile, <i>mean_logit_SSR_pval_j</i>, <i>mean_logit_SSR_y_j</i> and <i>mean_logit_SSR_t_j</i> are the averages of the logit-transformed <i>CS_survey_Q1</i> estimates for the replication probabilities for claims that fall into the same category in terms of p-value, time period and topic, as claim <i>j</i>, respectively.</p>
<i>CS_survey_Q1_AI</i>	Yes	<p>Forecast from survey responses as elicited in survey question Q1, using an alternative aggregation method. This score is computed similarly as in <i>CS_survey_Q1</i>, but with a rank sum score being the underlying metric for our surrogate score methods, instead of Brier score.</p> <p>The only difference to <i>CS_survey_Q1</i> is that:</p> $S'(SQ1_{i,j}, Y'_j) = w_{Y'_j} * \text{RankSum}(SQ1_{i,j}, Y'_j)$

		$+w_{1-Y'_j} * \text{RankSum}(SQ1_{i,j}, 1 - Y'_j).$ <p>As above, these scores are used to select the top performers up to date and perform mean aggregation over their forecasts. Rank-sum scoring rule can be found in Parry, Matthew. "Linear scoring rules for probabilistic binary classification." <i>Electronic Journal of Statistics</i> 10.1 (2016): 1596-1607.</p>
<i>CS_survey_Q1_A2</i>	Yes	<p>Forecast from survey responses as elicited in survey question Q1, using variance-weighted average survey responses. This score exploits that forecasters differ in the variance of their responses across the questions, and forecasters with a higher variance in their responses tend to provide better predictions.</p> $CS_survey_Q1_A2_j = \frac{\sum_i w_i SQ1_{i,j}}{\sum_i w_i}$ <p>The weight w_i denotes the variance of participant i over their responses to survey question Q1 in that batch. All responses are included.</p>
<i>CS_survey_Q3</i>	No	<p>Forecast from survey responses as elicited in survey question Q3. This score is optimized to forecast reproduction rather than replication, and is computed similarly as in <i>CS_survey_Q1_A1</i>, but with surrogate scores computed using elicited predictions for Q3 instead of Q1.</p>
<i>CS_survey_merged</i>	Yes	<p>Merged Q1-based and Q3-based surrogate scores. Survey question Q6 elicits if a claim is more or less likely than an average claim to be selected for replication. $CS_survey_merged_j$ is set to $CS_survey_Q1_A1_j$ if the average survey response to Q6 for this claim is equal or larger than the average response to Q6 across all claims in a round. Otherwise it is set to $CS_survey_Q3_j$.</p>
<i>CS_sp_Q1</i>	Yes	<p>This score is based on the ‘surprisingly popular’ score from Bayesian Truth Serum (BTS) method, and is calculated from survey questions Q1 and Q2. A proxy ground truth outcome is firstly identified using the a method similar to the ‘surprisingly popular’ method from BTS:</p> <p>For each claim j, we draw Y'_j from</p> $\text{Bernoulli}(2 * \text{Mean}_{\text{over}_i}(SQ1_{i,j}) - \text{Mean}_{\text{over}_i}(SQ2_{i,j}))$ <p>This random proxy will be used to compute a “proxy score” to evaluate each participant’s forecasts:</p> $ES_{i,j} = \mathbb{E}_{Y'_j} [\text{Brier}(SQ1_{i,j}, Y'_j)]$ <p>We then take the mean of $ES_{i,j}$ for each participant i that</p>

		<p>$MES_i = \text{Mean}_{\text{over } j}(ES_{i,j})$.</p> <p>On each claim, we select answers from the top (max(5, 10%*number of forecasts on the forecast question)) participants w.r.t. MES_i up to date and perform mean aggregation over their forecasts on that claim. For further information on BTS see Palley, A. B., & Soll, J. B. (2019). Extracting the Wisdom of Crowds When Information Is Shared. <i>Management Science</i>, 65(5), 2291-2309.</p>
CS_{sp_Q3}	No	Based on the ‘ surprisingly popular ’ score from BTS, and analogous to CS_{sp_Q1} ; but targeting reproduction using survey questions Q3 and Q4.
CS_{sp_merged}	No	Merged Q1/Q2-based and Q3/Q4-based ‘surprisingly popular’ scores. $CS_{sp_merged_j}$ is set to $CS_{sp_Q1_j}$ if the average response to Q6 for this claim is equal or larger than the average response to Q6 across all claims in a round. Otherwise it is set to $CS_{sp_Q3_j}$.
Other Scores		
$CS_{initial}$	No	Replication probability based on p-value , as estimated from previous replication projects (2). The p-value associated with the original claim is one of the strongest claim-related predictors of replication. $CS_{initial}$ is used for initial pricing of the markets.
CS_{meta}	No	<p>The meta score is the p-value dependent replication rate as extrapolated from past replication markets, moderated by journal-specific and time-specific predictions from the meta-markets.</p> $\text{logit}(CS_{meta_j}) = \text{logit}(CS_{initial_j}) + \Delta_j^T + \Delta_j^Y$ <p>The correction terms are calculated as</p>

		$\Delta_j^Y = \text{logit}(\text{meta_market_y}_j) - \text{logit}(\text{meta_market_all})$ $\Delta_j^T = \text{logit}(\text{meta_market_t}_j) - \text{logit}(\text{meta_market_all})$ <p>Here, meta_market_y_j and meta_market_t_j denote the estimates from the meta-markets for replication rates from publications of the same time period, and same topic, as claim j, respectively; meta_market_all denotes the overall replication rate in SCORE, as estimated with the corresponding question in the meta-markets.</p>
CS_survey_market	S YE	<p>Weighted average of surrogate score forecast ($\text{CS_survey_Q1_calibrated}$) and prediction market forecast ($\text{CS_market_calibrated}$). Weights are used to adjust e.g. for claims with low number of trades or survey participants.</p> $\text{CS_survey_market}_j = (w^S_j \text{CS_survey_Q1_calibrated} + w^M_j \text{CS_market_calibrated}) / (w^S_j + w^M_j)$ <p>with $w^S_j = \max(n^S_j/16, 1)$, and $w^M_j = \max(n^T_j/25, 1)$.</p> <p>The variables n^S_j, and n^T_j, denote the number of survey responses to survey question Q1 on claim j, and the number of trades on claim j on the claim-specific market. Our choice entails that once 25 trades, or 16 survey responses are reached, weights do not increase any longer. For claims with more than 25 trades and 16 survey responses, the survey estimate and the market estimate contribute with equal weights to this score.</p>

3.6 Amendments

3.6.1 Sep 1, 2019 (before start of data collection of Round 1)

- Time periods for Round 1 survey-based and market-based forecasting:

- Round 1 surveys set to: 2019-09-09, noon UTC – 2019-09-15, 23:59 UTC
- Round 1 markets set to: 2019-09-16, noon UTC – 2019-29-15, 23:59 UTC
- Subsequent rounds will follow in a 4 week cycle, with an extra week off for New Years'. If no other time off is taken, Round 10 would end on 2020-05-31
- Addition of two CS's. These scores will not be reported as part of the 10 reported scores.

<i>CS_survey_Q5_m</i> <i>ean</i>	No	Average response to survey question Q5 over all participants
<i>CS_market_QF</i>	No	Average response to all estimates provided on the market platform to the field " <i>Privately, I think the chance is about ...</i> ".

- Addition of two CS's. These scores will not be reported as part of the 10 reported scores.

3.6.2 Sep 9, 2019 (before start of data collection of Round 1)

- Change to rank sum score instead of Brier score as the underlying metric for CS_survey_Q1_calibrated, CS_survey_Q3, CS_survey_merged, CS_survey_market.

Recent theoretical findings suggest that the score based on rank-sum metrics (CS_survey_Q1_A1) can be expected to perform better than the score based on the Brier metric (CS_survey_Q1). Both methods are preregistered. A number of additional scores (CS_survey_Q1_calibrated, CS_survey_Q3, CS_survey_merged, CS_survey_market) were planned to be derived from CS_survey_Q1. Given the theoretical results, these scores will now be derived from CS_survey_Q1_A1. For the analysis of the meta-surveys, we will use both methods. Moreover, the rank-sum metric is used for payments.

3.6.3 Feb 1, 2020 (before start of data collection of Round 6)

Background: In Dec 2019, the T&E team proposed adjustments to the definition of replications as conducted for the SCORE project. Data-analytic replications that use a different, pre-existing dataset that is similar to the dataset used in the original study are now also part of the scope of replications used for validating scores. (The initial definitions distinguished between direct replications and data-analytic reproduction, the latter being conducted with the original data.) This shift in definition requires updating our forecasting approaches.

Prediction markets: For the prediction markets, contracts traded in Round 1-5 only pay if a claim was subject to direct replications. From round 6 onwards payouts will be made for contracts on claims that are evaluated through the broader replication definition (i.e. direct replication or data-analytic replication). This change affects the set of claims for which payouts will be made, but it does not affect the calculation of market-based scores provided to DARPA. Robustness checks will be used to detect changes in score characteristics due to this change.

Surveys: In Rounds 1-5 we used 7 questions as shown in Appendix B. Three of these questions (Q3, Q4, and Q6) are designed based on the assumption that participants distinguish between different types of replication and reproduction. An analysis of scores from Round 1-5 suggests that there is little evidence for this. Thus, in order to adjust the survey to the shift in the replication definitions, we simplify the survey by using the broader replication definition and omitting Q3, Q4, and Q6.

Scores: A number of scores are based on responses to Q3, Q4, and Q6: CS_survey_Q3, CS_survey_merged, CS_sp_Q3, and CS_sp_merged. One of these scores (CS_survey_merged) is among the 10 communicated to DARPA. The changes make these scores obsolete. Given the high correlation in Q1 and Q3-based scores, and to limit our analysis to scores that can be computed for

all the rounds of the project we will drop these scores for the full set of claims (i.e. from Round 1 onwards). CS_meta will replace CS_survey_merged as 10th score for DARPA.

The revised survey is shown in Appendix C; the new scores in Appendix D.

Modified hypothesis tests and additional robustness checks: In the hypothesis tests, tests involving variables derived from original Q3,4, and 6 will be dropped from the analysis. For tests involving Brier scores, the Brier scores will be calculated based on all (direct and data-analytic) replications. Changes are summarised in the Hp_characteristics.xlsx

Additionally, we test if mean forecasts changed, using a linear model of forecast as dependent variable, and Round (linear effect), post Round 5 (binary dummy); and field as independent variable; and whether Brier scores changed, using a linear model of Brier as dependent variable, and Round (linear effect), post Round 5 (binary dummy), and field as independent variable.

Other changes: Clarifications and changes to exclusion criteria.

Previously, our exclusion criteria for responses to the survey was: “Unless stated otherwise, we only use survey responses of participants who responded to at least 5 claim-level surveys.” This has been adjusted for clarity to: “Unless stated otherwise, survey related:

- *hypotheses* exclude participants who provided fewer than **5 claim surveys across all rounds**;
- *confidence scores including SSR* exclude participants who have completed fewer than **5 claimsurveys in that round**;
 - CS_survey_Q1_A2 also requires 5 completed claim surveys **in that batch** (due to the batch-wise calculation of variance).
- *prizes* further exclude participants who completed fewer than **9 claim surveys in that batch.**”

We have also removed exclusion criteria in score descriptions – see below.

3.6.4 April 11, 2020 (before start of data collection for the Round 6 markets)

Background: After the Round 6 surveys, the opening of the markets was delayed for additional IRB review. We here document changes in the schedule, and updates in the prediction market incentives from the new design used in round 6. Additionally, we change three scores (*CS_market_calibrated*, *CS_survey_Q1_calibrated* and *CS_meta*) by centring the corrections used in the calculations.

Schedule: Round 6 markets: 2020-04-13, noon UTC – 2019-04-27, 06:00 UTC

Incentives: The prize pool for the Round 6 markets is \$750 x # resolving claims. We expect about 125 claims to resolve under the new round 6 replication definitions.

Scores: The current approach to calculate *CS_market_calibrated*, *CS_survey_Q1_calibrated* and *CS_meta* ‘over-corrects’ these scores: The market prices for the individual claims, for instance, tend to be approximately 5% higher than what one would expect from the meta-markets (~50% vs. 45%). Applying the calibration as currently proposed (using three factors), we move the average forecast down to 35%, which is not justified. Instead of making small adjustments (~ 5%) based on the meta-markets we are calibrating nearly all claims down 15%.

We mitigate this by centring the correction. The revised scores are then defined as:

$\text{logit}(CS_market_calibrated_j)$

$$= \text{logit}(CS_market_smoothed_j) + \Delta^P_j + \Delta^T_j + \Delta^Y_j - \frac{\sum_{j=1}^{n_claims} \Delta_j^P}{n_claims} - \frac{\sum_{j=1}^{n_claims} \Delta_j^T}{n_claims} - \frac{\sum_{j=1}^{n_claims} \Delta_j^Y}{n_claims}$$

$\text{logit}(CS_survey_Q1_calibrated_j)$

$$= \text{logit}(CS_survey_Q1_AI_j) + \Delta'^P_j + \Delta'^Y_j + \Delta'^T_j - \frac{\sum_{j=1}^{n_claims} \Delta_j'^P}{n_claims} - \frac{\sum_{j=1}^{n_claims} \Delta_j'^T}{n_claims} - \frac{\sum_{j=1}^{n_claims} \Delta_j'^Y}{n_claims}$$

$$\text{logit}(CS_meta_j) = \text{logit}(CS_initial_j) + \Delta^T_j + \Delta^Y_j - \frac{\sum_{j=1}^{n_claims} \Delta_j^T}{n_claims} - \frac{\sum_{j=1}^{n_claims} \Delta_j^Y}{n_claims}$$

3.7 Appendices

3.7.1 Appendix A – Meta-questions

MSQ1) What will be the average replication rate in score?

MSQ2) What will be the average replication rate in economics?

MSQ3) What will be the average replication rate in political sciences?

MSQ4) What will be the average replication rate in psychology?

MSQ5) What will be the average replication rate in education research?

MSQ6) What will be the average replication rate in sociology and criminology?

MSQ7) What will be the average replication rate in marketing, management and related areas?

MSQ8) What will be the average replication rate for papers published in 2009/10?

MSQ9) What will be the average replication rate for papers published in 2011/12?

MSQ10) What will be the average replication rate for papers published in 2013/14?

MSQ11) What will be the average replication rate for papers published in 2015/16?

MSQ12) What will be the average replication rate for papers published in 2017/18?

3.7.2 Appendix B - Surrogate scoring questions

SQ1) What is the probability that a high-power **direct** replication of this study would find a statistically significant effect at the .05 level in the same direction as the original claim (0-100%)?

(Direct replications involve testing the original claim by gathering new data. 0 means that you think that a direct replication would never succeed, even by chance. 100 means that you think that a direct replication would never fail, even by chance.)

[slider 0 -100]

SQ2) What fraction of participants will give an estimate larger than 50% on Question 1?

[slider 0 -100]

SQ3) What is the probability that a high-power **data-analytic** replication of this study would find a statistically significant effect at the .05 level in the same direction as the original claim (0-100%)?

(Data-analytic replications involve testing the original claim using a similar, pre-existing dataset, for example the same economic indicator, but 5 years later. Again, 0 means that you think that a data-analytic replication would never succeed, even by chance. 100 means that you think that a data-analytic replication would never fail, even by chance.)

[slider 0 -100]

SQ4) What fraction of participants will give an estimate larger than 50% on Question 1?

[slider 0 -100]

SQ5) Considering the claim itself, without considering the specific implementation, how plausible is this claim (0-100)?

(0 means you think it cannot be true; 100 means you think it must be true.)

[slider 0 -100]

SQ6) Overall, the replication team will select about 100 claims for **direct** replication. Compared to average, how likely is this claim to be so chosen?

[Selection: Less likely; Average; More likely]

SQ7) Is there anything else we should know? For example, you may elaborate on your reasoning, report an error in the summary, note the claim is hard to understand, etc.

[Free text response]

3.7.3 Appendix C - Surrogate scoring questions Round 6-10

SQ1) What is the probability that a high-power replication of this study would find a statistically significant effect at the .05 level in the same direction as the original claim (0-100%)?

[slider 0 -100]

SQ2) What fraction of participants will give an estimate larger than 50% on Question 1?

[slider 0 -100]

SQ3) Considering the claim itself, without considering the specific implementation, how plausible is this claim (0-100)?

(0 means you think it cannot be true; 100 means you think it must be true.)

[slider 0 -100]

SQ4) Is there anything else we should know? For example, you may elaborate on your reasoning, report an error in the summary, note the claim is hard to understand, etc.

[Free text response]

3.7.4 Appendix D – New Scores

Name	Included in the 10 CSs	Short description
<i>Market-based scores</i>		
<i>CS_market_raw</i>	Yes	<p>Closing price of the markets on the binary claims. The market platform uses a logarithmic market scoring rule (LMSR) to update prices as shares are traded. After trade n in the shares of a CVD claim j, the market price for this share is updated from price $PR_j(n-1)$ to $PR_j(n)$. The score $CS_market_raw_j$ is the updated price after the final trade on claim j. This is analogous to what has been used in previous replication markets.</p>
<i>CS_market_smoothed</i>	Yes	<p>Smoothed market score. In previous replication markets 95% in the reduction of the forecasting errors occurs in the first 31 hours (56 trades) of the market (2). Afterwards there is little improvement in accuracy, and trading appears to be more noisy than informative. We aim to reduce the effect of noise by averaging the prices PR_j for a claim j over the second half of the trading period. To do so, we find the last trade k in the first half of the trading period and the last trade m in the second half of the trading period, and calculate</p> $CS_market_smoothed_j = \sum_{n=k..m} \frac{PR_j(n)(t_j(n+1) - t_j(n))}{t_j(m+1) - t_j(k)}$ <p>Here, $t_j(n)$ denotes the time of trade n on claim j; $t_j(m+1)$ denotes the closing time of the market. The time period used for averaging will be a subject of round-by-round review and might be adjusted depending on the distribution of trades over time.</p>
<i>CS_market_calibrated</i>	Yes	<p>Smoothed and calibrated market score. Previous replication markets have shown that a calibration of market prices can improve forecasts: p-values are not sufficiently priced in, and overall markets tend to be too optimistic (2). Forecasts from smoothed market prices can thus likely be improved by using information from the meta-markets, p-value, and past projects. We adjust the smoothed market score for each claim based on the CVD p-value, a topic-specific correction, and a time-specific correction. The correction will be applied in log-odds space to ensure the resulting scores remain between zero and one.</p>

		$\text{logit}(CS_market_calibrated_j)$ $= \text{logit}(CS_market_smoothed_j) + \Delta_j^P + \Delta_j^T + \Delta_j^Y$ <p>The correction terms are calculated as</p> $\Delta_j^P = \text{logit}(CS_initial_j) - \text{mean_logit_pval}_j$ $\Delta_j^Y = \text{logit}(meta_market_y_j) - \text{mean_logit_y}_j$ $\Delta_j^T = \text{logit}(meta_market_t_j) - \text{mean_logit_t}_j$ <p>Here, $meta_market_y_j$ and $meta_market_t_j$ denote the estimates from the meta-markets for replication rates from publications of the same time period, and same topic, as claim j, respectively. Additionally, $mean_logit_pval_j$, $mean_logit_y_j$, and $mean_logit_t_j$, are the averages of the logit-transformed market estimates for the replication probabilities for claims that fall into the same category in terms of p-value, time period, and topic, as claim j, respectively.</p>
Survey-based scores		
<i>CS_survey_Q1_mean</i>	No	Average response to survey question Q1 over all included participants
<i>CS_survey_Q1</i>	Yes	<p>Forecast from survey responses as elicited in survey question Q1, using the aggregation method with best track-record on similar data, the surrogate-score-aided aggregator (3). A surrogate score, using the Brier Score as the underlying metric, is computed and maintained for each survey participant to track their performance.</p> <p>To compute such score for each response of Q1, we first construct a (random) reference answer Y_j' for each claim j. Y_j' is a binary random variable following a Bernoulli distribution with parameter $\text{Mean}_{\text{over-}j}(SQI_{i,j})$.</p> <p>Our estimation algorithms will then compute two hyper-parameters w_1, w_0. Then for each response $SQI_{i,j}$, the surrogate score under our constructed reference Y_j' is</p>

		$S'(SQI_{i,j}, Y_j') = w_{Y_j'} * \text{Brier}(SQI_{i,j}, Y_j') + w_{1-Y_j'} * \text{Brier}(SQI_{i,j}, 1-Y_j').$ <p>We compute the expected surrogate score for $SQI_{i,j}$ as</p> $ES_{i,j} = \mathbb{E}_{Y_j' \sim \text{Bern}(\text{Mean}_{\text{over}_i}(SQI_{i,j}))} [S'(SQI_{i,j}, Y_j')]$ <p>For each participant i, we then take the mean of $ES_{i,j}$:</p> $MES_i = \text{Mean}_{\text{over}_j}(ES_{i,j})$ <p>On each claim, we select answers from the top (max(5, 10%*number of forecasts on the forecast question)) participants w.r.t. MES_i up to date and perform mean aggregation over their forecasts on that claim. For further information on the methodology see Liu, Y., & Chen, Y. (2018). Surrogate Scoring Rules and a Dominant Truth Serum. <i>arXiv preprint arXiv:1802.09158</i>.</p>
<p><i>CS_survey_Q1_calibrated</i></p>	<p>Yes</p>	<p>Calibrated CS_survey_Q1 scores. Forecasts from <i>CS_survey_Q1</i> may also likely be improved by using information from the meta-surveys, p-values, and past projects. We have:</p> $\text{logit}(CS_survey_Q1_calibrated_j)$ $= \text{logit}(CS_survey_Q1_AI_j) + \Delta^P_j + \Delta^Y_j + \Delta^T_j$ $\Delta^P_j = \text{logit}(CS_initial_j) - \text{mean_logit_SSR_pval}_j$ $\Delta^Y_j = \text{logit}(meta_survey_y_j) - \text{mean_logit_SSR_y}_j$ $\Delta^T_j = \text{logit}(meta_survey_t_j) - \text{mean_logit_SSR_t}_j$ <p>Here, $meta_survey_y_j$, $meta_survey_t_j$ denote the estimates from the meta-surveys for replication rates from publications of the same time period, and same topic, as claim j,</p>

		respectively. Meanwhile, $mean_logit_SSR_pval_j$, $mean_logit_SSR_y_j$ and $mean_logit_SSR_t_j$ are the averages of the logit-transformed CS_survey_Q1 estimates for the replication probabilities for claims that fall into the same category in terms of p-value, time period and topic, as claim j , respectively.
1	$CS_survey_Q1_A$	<p>Yes</p> <p>Forecast from survey responses as elicited in survey question Q1, using an alternative aggregation method. This score is computed similarly as in CS_survey_Q1, but with a rank sum score being the underlying metric for our surrogate score methods, instead of Brier score.</p> <p>The only difference to CS_survey_Q1 is that:</p> $S'(SQ1_{i,j}, Y'_j) = w_{Y'_j} * RankSum(SQ1_{i,j}, Y'_j) + w_{1-Y'_j} * RankSum(SQ1_{i,j}, 1 - Y'_j).$ <p>As above, these scores are used to select the top performers up to date and perform mean aggregation over their forecasts. Rank-sum scoring rule can be found in Parry, Matthew. "Linear scoring rules for probabilistic binary classification." <i>Electronic Journal of Statistics</i> 10.1 (2016): 1596-1607.</p>
2	$CS_survey_Q1_A$	<p>Yes</p> <p>Forecast from survey responses as elicited in survey question Q1, using variance-weighted average survey responses. This score exploits that forecasters differ in the variance of their responses across the questions, and forecasters with a higher variance in their responses tend to provide better predictions.</p> $CS_survey_Q1_A2_j = \frac{\sum_i w_i SQ1_{i,j}}{\sum_i w_i}$ <p>The weight w_i denotes the variance of participant i over their responses to survey question Q1 in that batch.</p>
	CS_survey_Q3	dro
d	CS_survey_merge	dro
	CS_sp_Q1	<p>Yes</p> <p>This score is based on the 'surprisingly popular' score from Bayesian Truth Serum (BTS) method, and is calculated from survey questions Q1 and Q2. A proxy ground truth outcome is firstly identified using the a method similar to the 'surprisingly</p>

		<p>popular' method from BTS:</p> <p>For each claim j, we draw Y_j' from Bernoulli ($2 * \text{Mean}_{\text{over}_i}(SQ1_{i,j}) - \text{Mean}_{\text{over}_i}(SQ2_{i,j})$)</p> <p>This random proxy will be used to compute a “proxy score” to evaluate each participant’s forecasts:</p> $ES_{i,j} = \mathbb{E}_{Y_j'}[\text{Brier}(SQ1_{i,j}, Y_j')]$ <p>We then take the mean of $ES_{i,j}$ for each participant i that $MES_i = \text{Mean}_{\text{over}_j}(ES_{i,j})$.</p> <p>On each claim, we select answers from the top ($\max(5, 10\% * \text{number of forecasts on the forecast question})$) participants w.r.t. MES_i up to date and perform mean aggregation over their forecasts on that claim. For further information on BTS see Palley, A. B., & Soll, J. B. (2019). Extracting the Wisdom of Crowds When Information Is Shared. <i>Management Science</i>, 65(5), 2291-2309.</p>
<i>CS_sp_Q3</i>	pped dro	
<i>CS_sp_merged</i>	pped dro	
Other Scores		
<i>CS_initial</i>	No	<p>Replication probability based on p-value, as estimated from previous replication projects (2). The p-value associated with the original claim is one of the strongest claim-related predictors of replication. <i>CS_initial</i> is used for initial pricing of the markets.</p>
<i>CS_meta</i>	NE W: YES	<p>The meta score is the p-value dependent replication rate as extrapolated from past replication markets, moderated by journal-specific and time-specific predictions from the meta-markets.</p> $\text{logit}(CS_{\text{meta}_j}) = \text{logit}(CS_{\text{initial}_j}) + \Delta^T_j + \Delta^Y_j$

		<p>The correction terms are calculated as</p> $\Delta_j^Y = \text{logit}(\text{meta_market_y}_j) - \text{logit}(\text{meta_market_all})$ $\Delta_j^T = \text{logit}(\text{meta_market_t}_j) - \text{logit}(\text{meta_market_all})$ <p>Here, meta_market_y_j and meta_market_t_j denote the estimates from the meta-markets for replication rates from publications of the same time period, and same topic, as claim j, respectively; meta_market_all denotes the overall replication rate in SCORE, as estimated with the corresponding question in the meta-markets.</p>
<p>t</p> <p><i>CS_survey_marke</i></p>	<p>S</p> <p>YE</p>	<p>Weighted average of surrogate score forecast (<i>CS_survey_Q1_calibrated</i>) and prediction market forecast (<i>CS_market_calibrated</i>). Weights are used to adjust e.g. for claims with low number of trades or survey participants.</p> $CS_survey_market_j = (w_j^S CS_survey_Q1_calibrated + w_j^M CS_market_calibrated) / (w_j^S + w_j^M)$ <p>with $w_j^S = \max(n_j^S/16, 1)$, and $w_j^M = \max(n_j^T/25, 1)$.</p> <p>The variables n_j^S, and n_j^T, denote the number of survey responses to survey question Q1 on claim j, and the number of trades on claim j on the claim-specific market. Our choice entails that once 25 trades, or 16 survey responses are reached, weights do not increase any longer. For claims with more than 25 trades and 16 survey responses, the survey estimate and the market estimate contribute with equal weights to this score.</p>

3.8 COVID19 – Claims pre-registration documentation for SCORE TA2/Team KeyW

3.8.1 Context

With its SCORE programme, the Defense Sciences Office (DSO) at the Defense Advanced Research Projects Agency (DARPA) is funding innovative research projects for the development and deployment of tools to assign Confidence Scores (CSs) to different kinds of Social and Behavioral Science (SBS) research results and claims. CSs are quantitative measures that should enable someone to understand the degree to which a particular claim or result is likely to be reproducible and/or replicable. In the first phase of the SCORE programme, a total of 3000 claims from a broad range of academic disciplines were scored, with only a small fraction of claims being validated through direct replication and data-analytic replications conducted by a separate team (TA1).

In an additional round, conducted in August and September 2020 we score 100 COVID-19 related claims from the SBS literature, 20-50% of which will be validated (largely through data-analytic replications) by TA1. Our team will use surrogate scoring and prediction market approaches to elicit CSs from participants. We will not receive information about which claims will be selected for reproduction or replication and will not receive claim-specific detail on replications. The outcome of the TA1 replications will be used to evaluate the scores provided by our team.

3.8.1.1 Methodology

Overview. As done for the original 3000 SCORE claims, we use two methodologies to elicit forecasts on the replicability of claims: decision markets and surrogate scoring. Decision markets are mechanisms related to prediction markets that are modified to provide incentivized forecasts despite

the low rate of resolution. Surrogate scoring is a survey-based incentivized elicitation method that does not require access to a ‘ground truth’ to determine incentives. Of 9 submitted scores, 2 will be based on the decision markets, 5 will be based on the surrogate scoring surveys, 1 will be based on p-value, and 1 will use information from both approaches. A description of these scores is given in Section 4. Forecasts on the replicability of the 100 claims are elicited in an additional single round within the SCORE project, starting in August 2020. The round will start with surveys for the surrogate scoring method. Each participant will be assigned to a batch with survey questions on 10 of the 100 claims in that round. Once participants complete this first batch they can choose to complete a different batch of claims. The surveys will be accessible for one week. Once the survey closes, markets on the replicability of the 100 claims will open for four weeks. In addition to the participants from the original SCORE project, we will recruit participants through a variety of community mailing lists and social media; survey and markets can be accessed through the website www.replicationmarkets.com. Once registered, participants new to SCORE are asked to complete compulsory and non-compulsory intake surveys.

Surrogate scoring. Surrogate scoring will be used to elicit forecasts on the outcome of replications for the 100 COVID related claims. In addition, the survey questions will address predictions on the beliefs of other forecasters, the plausibility of claims, and errors in the claim data (see [Appendix A](#)). Survey responses will be elicited for batches of 10 claims (see above). Incentives will be provided only for forecasts on the outcome of replications (SQ1). The surrogate scoring methodology will be used to identify and incentivize the expected top forecasters in each round for each batch. The surveys will be open for one week at the beginning of each round. The aggregation of survey responses into forecasts is described in Section 4.

Prediction markets. Once the surrogate scoring survey is closed, the market for the replicability of the 100 individual claims in the round will open. The main market-based mechanism

to elicit forecasts for the outcome of the individual replications will be through the trading of ‘shares’ on binary (yes/no) questions. For each claim there will be ‘Yes’ shares and ‘No’ shares. ‘Yes’ shares will pay 1 point if the underlying claim is evaluated through replication, and the replication yields a statistically significant finding in the direction of the original claim. ‘No’ shares will pay 1 point if the underlying claim is evaluated through replication, and the replication does not yield a statistically significant finding in the direction of the original claim. If a claim is not selected for replication, both ‘Yes’ shares and ‘No’ shares will pay out 0 points. This approach is equivalent to using decision markets with simple stochastic decision rules. Trading will be facilitated through a market maker implementing a logarithmic market scoring rule (LMSR) with base 2 and a liquidity parameter of $b = 100$. Initial decision market prices will be informed by p-values and the relation between p-value and replication rates as observed in previous replication markets. Each claim will be assigned into one of the following three categories (“ $p \leq 0.001$ ”, “ $0.001 < p \leq 0.01$ ”, “ $p > 0.01$ ”). Starting prices for these three categories will be (0.8, 0.4, 0.3). Participants initially receive 100 points, and can obtain a bonus of up to 30% depending on ongoing contributions to the market and survey. Halfway through participants will receive another 40 points (plus up to 30% bonus). The market will be open for 4 weeks. Points that are not invested by the time the market closes will be lost. Forecasts will be generated by smoothing the market prices over the last week of the trading period (see Section 5, *CS_market_smoothed* for details).

Prizes. The total prize pool will be split into one part dedicated to the prediction markets, and one part dedicated to surrogate scoring. Replication results are expected to be released at the end of the project (currently anticipated for 30-NOV-2020). The prize pool for the prediction markets will thus be paid out at the end of the project. Prizes for surrogate scoring will be paid out after the markets close. This ensures that information on performance in surrogate scoring does not affect investments in the market. The prize pool for surrogate scoring will be allocated into prize pools of USD 160 for

the individual batches. The top four participants for any given batch (as determined by their average score over the round; see Section 5 for details) will receive fixed prizes of USD 80, USD 40, USD 20, and USD 20. The prize pool for the prediction market will be USD 9,000. This prize pool will be allocated proportionally to the points that participants earn from their investments. Participants in the prediction markets do not receive any payouts for points that have not been invested into forecasts.

3.8.2 Pre-registration for the statistical analysis of the regular markets and surveys

Unless stated otherwise, we only use survey responses of participants who responded to at least 5 claim-level surveys. Unless otherwise specified, we interpret the threshold of $p < 0.005$ as identifying statistical significance and the threshold of $p < 0.05$ as identifying suggestive evidence; all the tests in this pre-analysis plan are two-sided tests.

3.8.2.1 Primary Questions

1. **How well calibrated are the CSs (see Section 5; all scores are used)? Do they overestimate or underestimate actual replication rates?**

Tests: Using only claims with attempted replications, we conduct a paired t-test between binary outcome data (1 = replicated, 0 = not replicated) and each of the corresponding computed CSs.

Priority for reporting will be given to survey means CS_survey_Q1_mean, the surrogate score CS_survey_Q1, and the raw market forecast CS_market_raw. The results for the other scores will be treated as *secondary*.

2. **How do the scores differ from each other in terms of forecasting performance. Which CS(s) perform best?**

Tests: Brier scores are calculated for each of the CSs with observed **replication outcome**. We then use pairwise paired t-tests of the Brier scores to test for differences in the mean. To control the false discovery rate, we use the Benjamini – Hochberg correction.

We expect that the final market prices are improved over the initial market prices, and that smoothing improves forecasts. For the survey-based forecasts, we expect that all surrogate scoring methods are improved over simple survey averages. It is plausible that merging market forecasts with survey-based forecasts as outlined in Section 5 leads to further improvement over both the smoothed market scores and the surrogate scores. Priority for reporting will thus be given to comparisons between **market based scores** (initial score *CS_initial* vs. raw market price *CS_market_raw*; raw market price *CS_market_raw* vs smoothed market price *CS_market_smoothed*; smoothed market price *CS_market_smoothed* vs merged survey and market score *CS_survey_market*), and between **survey based scores** (survey mean *CS_survey_Q1_mean* vs surrogate scores *CS_survey_Q1*, surrogate scores *CS_survey_Q1* vs merged survey and market score *CS_survey_market*). All other comparisons will be treated as secondary/exploratory hypotheses.

3. Is there a relation between the p-value of the original study and the rate of replication?

Tests: We use a linear model with replication outcome as dependent variable, and original study p-value category as independent variable. We use the same categorization as used for the initial market prices.

3.8.2.2 Secondary Hypotheses

4. Is career stage associated with forecasting accuracy?

Tests: For each participant, and each claim validated through replication, we calculate the Brier score for the response to survey Question Q1. We then conduct a linear regression of the Brier score and career stage as obtained in the demographic questionnaire. Clustering the standard errors of participants is used.

5. **Is the variance of a participant's survey responses correlated with their overall forecasting accuracy?**

Tests: For each participant, we calculate the variance over all responses given to Q1, and the mean Brier score using response to Q1 and observed replication outcomes. We use Pearson's correlation coefficient to test for a correlation between these two variables.

6. **Are number of trades correlated with the Brier score of final market price?**

Tests: We use Spearman's correlation coefficient to test for a correlation between the number of trades in the market, and the Brier score for forecasts from *CS_market_raw* calculated based on the replication outcomes.

7. **Is traded volume (in terms of shares traded) correlated with the Brier score for the market forecast**

Tests: We use Spearman's correlation coefficient to test for a correlation between the number of traded shares for a claim, and the Brier score for forecasts from *CS_market_raw* calculated based on the replication outcomes.

Disclaimer:. Changes will be documented in amendments to this document. We reserve the right to adjust analyses and to test additional hypotheses. This will be indicated in the resulting documentation.

3.8.2.3 Additional descriptive analyses

We calculate Pearson correlation coefficients for each pair of CSs. We calculate for each score the mean absolute forecasting error, mean Brier score (including decompositions), AUC, and the R^2 , intercept and coefficient of a linear model between CSs and the observed replication outcomes. *The scores for all 100 claims are included for these tests.*

3.8.3 Scores

The table below summarizes the scores calculated by our team, and indicates which ones are reported to DARPA. Key variables include:

- From the **regular (claim-level) markets**, variables $t_j(n)$ and $PR_j(n)$, denote the time of trade number n in claim j , and the updated price after this trade.
- From the **claim-level surveys**, $SQ1_{i,j}$, $SQ2_{i,j}$, ..., denote the response of participant i to Question Q1, Q2, ..., in the surrogate scoring survey for claim j .

Table 3-2: COVID19 Confidence Score Descriptions

• Name	Included in the 10 CSs	Short description
<i>Market-based scores</i>		
<i>CS_market_raw</i>	Yes	<p>Closing price of the markets on the binary claims. The market platform uses a logarithmic market scoring rule (LMSR) to update prices as shares are traded. After trade n in the shares of a CVD claim j, the market price for this share is updated from price $PR_j(n-1)$ to $PR_j(n)$. The score $CS_market_raw_j$ is the updated price after the final trade on claim j. This is analogous to what has been used in previous replication markets.</p>
<i>CS_market_smoothed</i>	Yes	<p>Smoothed market score. In previous replication markets 95% in the reduction of the forecasting errors occurs in the first 31 hours (56 trades) of the market (2). Afterwards there is little improvement in accuracy, and trading appears to be more noisy than informative. We aim to reduce the effect of noise by averaging the prices PR_j for a claim j over the last week of the trading period. To do so, we find the last trade k in the first three weeks of the trading period and the last trade m in the last week of the trading period, and calculate</p> $CS_market_smoothed_j = \sum_{n=k..m} \frac{PR_j(n)(t_j(n+1) - t_j(n))}{t_j(m+1) - t_j(k)}$ <p>Here, $t_j(n)$ denotes the time of trade n on claim j; $t_j(m+1)$ denotes the closing time of the market.</p>

Survey-based scores		
<i>CS_survey_Q1_mean</i>	Yes	Average response to survey question Q1 over all included participants
<i>CS_survey_Q1</i>	Yes	<p>Forecast from survey responses as elicited in survey question Q1, using the aggregation method with best track-record on similar data, the surrogate-score-aided aggregator (3). A surrogate score, using the Brier Score as the underlying metric, is computed and maintained for each survey participant to track their performance.</p> <p>To compute such score for each response of Q1, we first construct a (random) reference answer Y_j' for each claim j. Y_j' is a binary random variable following a Bernoulli distribution with parameter $\text{Mean}_{\text{over}_j}(SQ1_{i,j})$.</p> <p>Our estimation algorithms will then compute two hyper-parameters w_1, w_0. Then for each response $SQ1_{i,j}$, the surrogate score under our constructed reference Y_j' is</p> $S'(SQ1_{i,j}, Y_j') = w_{Y_j'} * \text{Brier}(SQ1_{i,j}, Y_j') + w_{1-Y_j'} * \text{Brier}(SQ1_{i,j}, 1-Y_j').$ <p>We compute the expected surrogate score for $SQ1_{i,j}$ as</p> $ES_{i,j} = \mathbb{E}_{Y_j' \sim \text{Bern}(\text{Mean}_{\text{over}_i}(SQ1_{i,j}))} [S'(SQ1_{i,j}, Y_j')]$ <p>For each participant i, we then take the mean of $ES_{i,j}$:</p> $MES_i = \text{Mean}_{\text{over}_j}(ES_{i,j})$ <p>On each claim, we select answers from the top ($\max(5, 10\% * \text{number of forecasts on the forecast question})$) participants w.r.t. MES_i up to date and perform mean aggregation over their forecasts on that claim. For further information on the methodology see Liu, Y., & Chen, Y. (2018). Surrogate Scoring Rules and a Dominant Truth Serum. <i>arXiv preprint arXiv:1802.09158</i>.</p>
<i>CS_survey_Q1_AI</i>	Yes	<p>Forecast from survey responses as elicited in survey question Q1, using an alternative aggregation method. This score is computed similarly as in CS_survey_Q1, but with the rank sum score being the underlying metric for our surrogate score methods, instead of Brier score.</p> <p>The only difference to CS_survey_Q1 is that:</p> $S'(SQ1_{i,j}, Y_j') = w_{Y_j'} * \text{RankSum}(SQ1_{i,j}, Y_j') + w_{1-Y_j'} * \text{RankSum}(SQ1_{i,j}, 1 - Y_j').$ <p>As above, these scores are used to select the top performers up to date and perform mean aggregation over their forecasts. Rank-sum scoring rule can be found in Parry, Matthew. "Linear scoring rules for probabilistic binary</p>

		classification." <i>Electronic Journal of Statistics</i> 10.1 (2016): 1596-1607.
<i>CS_survey_Q1_A2</i>	Yes	<p>Forecast from survey responses as elicited in survey question Q1, using variance-weighted average survey responses. This score exploits that forecasters differ in the variance of their responses across the questions, and forecasters with a higher variance in their responses tend to provide better predictions.</p> $CS_survey_Q1_A2_j = \frac{\sum_i w_i SQ1_{i,j}}{\sum_i w_i}$ <p>The weight w_i denotes the variance of participant i over their responses to survey question Q1.</p>
<i>CS_sp_Q1</i>	Yes	<p>This score is based on the ‘surprisingly popular’ score from Bayesian Truth Serum (BTS) method, and is calculated from survey questions Q1 and Q2. A proxy ground truth outcome is firstly identified using the a method similar to the ‘surprisingly popular’ method from BTS:</p> <p>For each claim j, we draw Y_j' from</p> <p>Bernoulli ($2 * \text{Mean}_{\text{over}_i}(SQ1_{i,j}) - \text{Mean}_{\text{over}_i}(SQ2_{i,j})$)</p> <p>This random proxy will be used to compute a “proxy score” to evaluate each participant’s forecasts:</p> $ES_{i,j} = \mathbb{E}_{Y_j'} [\text{Brier}(SQ1_{i,j}, Y_j')]$ <p>We then take the mean of $ES_{i,j}$ for each participant i that</p> $MES_i = \text{Mean}_{\text{over}_j}(ES_{i,j}).$ <p>On each claim, we select answers from the top ($\max(5, 10\% * \text{number of forecasts on the forecast question})$) participants w.r.t. MES_i up to date and perform mean aggregation over their forecasts on that claim. For further information on BTS see Palley, A. B., & Soll, J. B. (2019). Extracting the Wisdom of Crowds When Information Is Shared. <i>Management Science</i>, 65(5), 2291-2309.</p>
Other Scores		
<i>CS_initial</i>	S YE	<p>Replication probability based on p-value, as estimated from previous replication projects. The p-value associated with the original claim is one of the strongest claim-related predictors of replication. <i>CS_initial</i> is used for initial pricing of the markets.</p>
<i>CS_survey_market</i>	S YE	<p>Weighted average of surrogate score forecast (<i>CS_survey_Q1</i>) and prediction market forecast (<i>CS_market_smoothed</i>). Weights are used to adjust e.g. for</p>

		<p>claims with low number of trades or survey participants.</p> $CS_survey_market_j = (w^S_j \cdot CS_survey_Q1 + w^M_j \cdot CS_market_smoothed) / (w^S_j + w^M_j)$ <p>with $w^S_j = \max(n^S_j/16, 1)$, and $w^M_j = \max(n^T_j/25, 1)$.</p> <p>The variables n^S_j, and n^T_j, denote the number of survey responses to survey question Q1 on claim j, and the number of trades on claim j on the claim-specific market. Our choice entails that once 25 trades, or 16 survey responses are reached, weights do not increase any longer. For claims with more than 25 trades and 16 survey responses, the survey estimate and the market estimate contribute with equal weights to this score.</p>
--	--	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

3.8.4 Appendices

3.8.4.1 Appendix A - Surrogate scoring questions

SQ1) What is the probability that a high-power replication of this study would find a statistically significant effect at the .05 level in the same direction as the original claim (0-100%)?

[slider 0 -100]

SQ2) What fraction of participants will give an estimate larger than 50% on Question 1?

[slider 0 -100]

SQ3) Considering the claim itself, without considering the specific implementation, how plausible is this claim (0-100)?

(0 means you think it cannot be true; 100 means you think it must be true.)

[slider 0 -100]

SQ4) Is there anything else we should know? For example, you may elaborate on your reasoning, report an error in the summary, note the claim is hard to understand, etc.

[Free text response]

3.9 References

1. DARPA. DARPA SCORE Internal Documentation. 2019.
2. Gordon M, Viganola D, Dreber A, Johannesson M, Pfeiffer T. Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. PLOS ONE. 2021 Apr 14;16(4):e0248780.
3. Liu Y, Wang J, Chen Y. Surrogate Scoring Rules. ACM Conf Econ Comput 2020 Jul; Available from: <http://arxiv.org/abs/1802.09158>

4 Chapter 4: Are replication rates the same across academic fields?

Community forecasts from the DARPA SCORE program

The paper “Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE program” published in 2020 in the journal “Royal Society Open Science” is found in this chapter. This paper contains the results of the first round of forecasts elicited as part of the SCORE project (as described in Appendix 1). The methodology was heavily informed by previous research, as outlined in Chapter 2. The “meta-forecasts” presented in this paper will also be used to inform Confidence Scores (see Chapter 3 for full details about confidence scores). My role in this paper was the experimental and statistical design, conducting the analysis and drafting the manuscript (including the visualisations).

The reference for this paper is: Gordon M, Viganola D, Bishop M, Chen Y, Dreber A, Goldfedder B, Holzmeister F, Johannesson M, Liu Y, Twardy C, Wang J. Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. Royal Society open science. 2020 Jul 22.

To align with the formatting and referencing style of this thesis, there are some changes in formatting and referencing style of the published paper

Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE program

4.1 Abstract

The DARPA program “Systematizing Confidence in Open Research and Evidence” (SCORE) aims to generate confidence scores for a large number of research claims from empirical studies in the social and behavioral sciences. The confidence scores will provide a quantitative assessment of how likely a claim will hold up in an independent replication. To create the scores, we follow earlier approaches and use prediction markets and surveys to forecast replication outcomes. Based on an initial set of forecasts for the overall replication rate in SCORE and its dependence on the academic discipline and the time of publication, we show that participants expect replication rates to increase over time. Moreover, they expect replication rates to differ between fields, with the highest replication rate in economics (average survey response 58%), and the lowest in Psychology and in Education (average survey response of 42% for both fields). These results reveal insights into the academic community’s views of the replication crisis, including for research fields for which no large-scale replication studies have been undertaken yet.

4.2 Introduction

Replication has long been established as a key practice in scientific research (1,2). It plays a critical role in controlling the impact of sampling error, questionable research practices, publication bias, and fraud (1,2). An increase in the effort to replicate studies has been argued to help establishing credibility within a field (3). Moreover, replications allow to test if results generalize to

a different or larger population, and help to verify the underlying theory (2,4,5) and its scope (6).

Despite the importance of replications, there are practical and resource-related constraints that limit the extent to which replications are conducted (7).

Previous studies have shown that information about replication outcomes can be elicited from the research community (8–11). This suggests that forecasting the outcomes of hypothetical replications can help assessing replication probabilities without requiring the resources for actually conducting replications. The DARPA program “Systematizing Confidence in Open Research and Evidence” (SCORE) follows this approach to generate confidence scores for thousands of research claims from empirical studies in the social and behavioral sciences. These confidence scores will provide a quantitative assessment of how likely a claim will hold up in an independent replication. A small subset of the claims (about 5%) will eventually be assessed through replication, and the replication outcomes will be used to evaluate the accuracy of the confidence scores. The research claims are sampled from studies published during a 10-year period (2009–2018) across 60 journals from a number of academic disciplines.

To generate confidence scores for the DARPA SCORE program (12) we follow the template of past forecasting studies (8–11) and use surveys and prediction markets. In the surveys, participants recruited from the relevant research communities are asked to provide their estimates for the probability that a claim will hold up in a replication. Replication here refers to either direct replication (i.e., same data collection process and analysis on a different sample) or data-analytic replication (i.e., same analysis on a similar but independent dataset; see (13,14)). Successful replication is defined as an effect that is in the same direction as the original effect, and statistically significant at $p < 0.05$. While replication projects such as SCORE typically provide additional and non-binary characteristics of the replication results, such a binary definition is well-suited for elicitation of forecasts through prediction markets and surveys. In the prediction markets (15–17),

participants trade contracts with payoffs tied to the outcome of replications and thereby generate prices that provide quantitative forecasts of the replication results (8–11). Because of the large number of claims to be assessed, forecasting takes place in monthly rounds from mid-2019 to mid-2020. In each monthly round, about 300 claims are assessed, resulting in about 3,000 claims assessed by mid-2020.

Before we started collecting claim-specific forecasts, we collected an initial set of surveys and market forecasts for the overall replication rate in SCORE and its dependence on the academic discipline and the time of publication. These forecasts allow us to test whether participants at the beginning of the SCORE project expect field-specific and time-dependent variation of the replication rates. In this paper, we present an analysis of the data from this initial round of forecasting. Whether and to which extent these meta-forecasts are correct will be explored once the replications have actually been conducted.

4.3 Methods

The surveys and prediction markets to forecast time- and discipline-specific replication rates were open for one week each (August 12–18, 2019 and August 19–25, 2019), with the market starting after the survey closed. We asked participants to forecast the overall SCORE replication rate, the replication rate in 5 non-overlapping 2-year periods (2009/10, 2011/12, 2013/14, 2015/16, and 2017/18), and in 6 discipline clusters (Economics, Political Science, Psychology, Education, Sociology and Criminology, and Marketing, Management and Related Areas). The discipline clusters are defined through journals (see Supplementary Table 1). Two of these clusters, namely Sociology and Criminology, and Marketing, Management and Related Areas, are heterogeneous in terms of research fields and are comprised of fields with a small number of journals sampled in SCORE. We combined those fields into clusters to meet a minimal number of journals per cluster.

To elicit forecasts, we used the same wording in the survey and the prediction market (see Supplementary Table 2). Further information, including the definition of what constitutes a successful replication, and the targeted power of the replications, were available in the online instructional material.

The surveys were incentivized using a peer assessment method which employs a Surrogate Scoring Rule. Surrogate scoring rules (18) provide an unbiased estimate of strictly proper scoring rules (19) and are dominantly truthful in eliciting probabilistic predictions. When scoring an agent's predictions, surrogate scoring rules first construct a noisy prediction of the event's outcome from other agents' reports (noisy 'ground truth'). The second step estimates the bias of this noisy 'ground truth' using all elicited predictions across all users on all forecasting questions. Then, surrogate scoring rules compute a de-biased version of strictly proper scoring rules. Consequently, to maximize one's expected surrogate score, it is always a dominant strategy to report truthfully, due to the incentive property of the strictly proper scoring rules. The total prize pool of \$960 was allocated in prizes of \$80 for the top 6 participants, \$40 for participants ranked 7–12, and \$20 for participants ranked 13–24.

In the prediction markets, the participants traded contracts that pay out points proportional to the actual replication rate in each time period and discipline cluster. The total prize pool of \$4,320 will be allocated once the replications are completed in proportion of the points earned from the contracts. To facilitate trading we used a market maker implementing a logarithmic market scoring rule (16) with base 2 and a liquidity parameter of $b = 100$. Participants received an initial endowment of 100 points to trade with. Findings from previous replication-focused prediction markets (8–11) showed that most of the information elicited in these markets gets typically priced in soon after the markets opened; after about 2 days of trading, prices tend to fluctuate around equilibrium prices. To minimize the impact of noisy price fluctuations on our forecasts, we use a

pre-registered time-weighted average of the prices from the second half of the trading period as market forecasts.

Participants were recruited through a number of mailing lists, Twitter, and blog posts. As of the start of the initial surveys, 478 forecasters had signed up to participate and expressed informed consent, and 226 subsequently completed the initial forecasting survey. Most of these participants (80%) were from academia, with smaller groups from the private sector, non-profit organizations, and the public sector (11%, 4%, and 4%, respectively). The majority (69%) of participants are in early (e.g., undergraduate, graduate student, postdoc, assistant professor) or mid (e.g., senior research fellow) career stages. Those at senior career stages (e.g., full or emeritus professor) made up 7.5% of the survey takers. Self-reported interests of our participants were dominated by two fields: 99 survey respondents (44%) indicated that they were interested in Economics, and 126 (56%) in Psychology. Marketing, Management and Related Areas, the next largest field, was selected by 45 participants (20%), followed by Political Science, Education, and Sociology and Criminology with 24 (11%), 23 (10%), and 22 (10%) responses, respectively. About half of all participants (122) only gave one field of interest, 65 (29%) respondents indicated two fields of interest, and 38 (17%) reported three or more fields of interest. 48% of the survey participants indicated that they had pre-registered at least one study before, and 34% have been involved in a replication study. 37% of survey takers indicated that they had participated in a prediction market prior to SCORE. 217 forecasters made at least one trade in the prediction markets. Given that 64% of these market traders also completed the survey, the participants in the market had a demographic composition similar to those who completed the survey.

The experimental design and statistical analyses were pre-registered on OSF, before the initial surveys and markets were opened. The data collected in the initial forecasting round allow us to analyse four of eight hypotheses provided in the pre-registration: (1) whether forecasted

replication rates differ between fields of research; (2) whether forecasted replication rates differ between time periods; (3) whether topic-specific forecasts depend on the forecasters' field of research; and (4) whether survey-based aggregated forecasts and market-based forecasts are correlated. For all analyses we use survey responses; for analysis 3 we additionally use responses to a demographic survey that included a question on academic interests; and for analysis 4 we additionally use the prediction market data. All these analyses were conducted as pre-registered. The remaining four pre-registered analyses require data that become available once the forecasting for the individual research claims and the replications are completed. For statistical tests, we interpret the threshold of $p < 0.005$ as identifying statistical significance, and the threshold of $p < 0.05$ as identifying suggestive evidence (20). Pre-registration document, data, codebook and scripts are available at the dryad repository (<https://doi.org/10.5061/dryad.pg4f4qrk5>).

4.4 Results

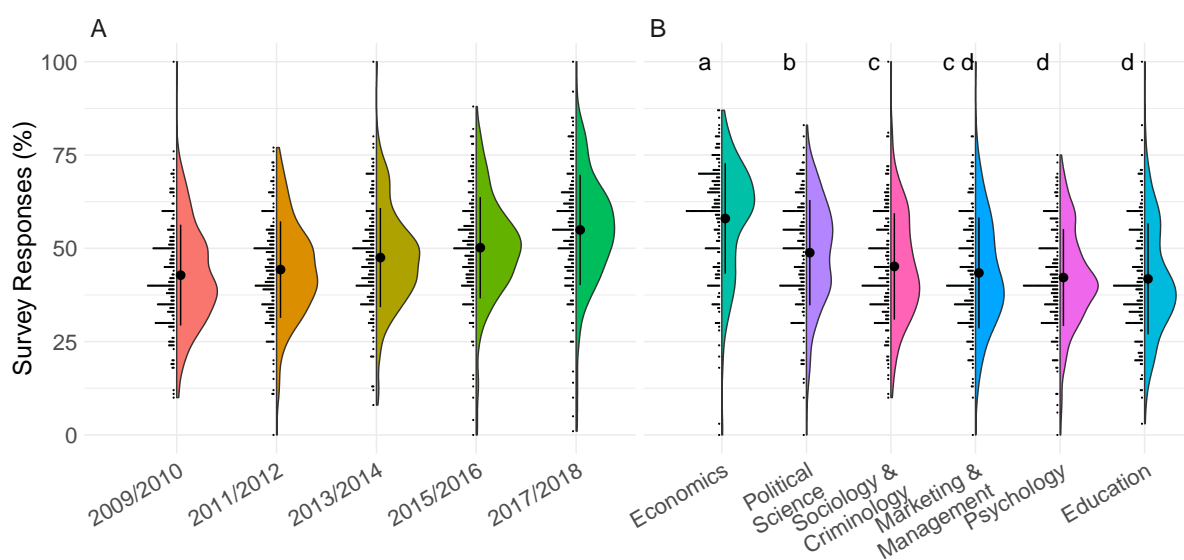
Summary statistics for the surveys and markets are provided in Table 1. The mean of the survey responses for the overall replication rate forecasts was 49%, which is close to the replication rate across previous replication projects (8,9,21,22). The survey results provide evidence that the participants expect the replications rates to differ across the two-year time periods (one-way repeated measures ANOVA; $F(4,900) = 130.5$, $p < 0.0001$). Replication rates were expected to increase over time from 43% in 2009/10 to 55% in 2017/18 (see Figure 1A). Using pairwise paired t -tests, with a Benjamini-Hochberg correction to control for the false discovery rate, all year bands were forecasted to have different replication rates significant at the 0.0001 level, except for the comparison between 2009/10 and 2011/12, where suggestive evidence is found for differences in replication rate expectations ($p = 0.0066$; see Supplementary Table 4 for all p -values and test statistics).

Table 4-1. Descriptive statistics of survey and market forecasts. Final price refers to the price at market closing, whereas smoothed price is a weighted average of the market prices, designed to reduce effects of noise near the end of the market. The aggregation methods are all highly correlated, with the Pearson's correlation coefficient between aggregation methods as follows: Smoothed Price and SSR Brier 0.963, Smoothed Price and SSR Rank 0.935, Smoothed Price and Survey Mean 0.957, SSR Brier and SSR Rank 0.924, SSR Brier and Survey Mean 0.942, SSR Rank and Survey Mean 0.979. All correlations are statistically significant at $p < 0.0001$ ($df=10$).

Overall Replication Rate in...	Smoothed Price	Final Price	Distinct Traders	Number of Trades	Survey Mean	SSR Rank	SSR Brier
Economics	0.57	0.58	156	235	0.58	0.65	0.36
Political Science	0.45	0.46	115	158	0.49	0.55	0.29
Psychology	0.41	0.45	165	226	0.42	0.39	0.26
Education	0.39	0.45	126	176	0.42	0.38	0.24
Sociology & Criminology	0.44	0.43	106	133	0.45	0.45	0.29
Marketing & Management	0.40	0.36	124	161	0.43	0.41	0.25
2009/10	0.41	0.40	74	107	0.43	0.41	0.25
2011/12	0.43	0.42	60	77	0.44	0.44	0.26
2013/14	0.45	0.44	64	80	0.48	0.49	0.27
2015/16	0.45	0.46	98	147	0.50	0.53	0.30
2017/18	0.50	0.49	82	104	0.55	0.58	0.32
all claims in SCORE	0.47	0.48	83	125	0.49	0.49	0.31

The survey responses also show that participants expected replication rates to differ between topics. Using responses from the survey, a one-way repeated measures ANOVA finds a joint significant effect of topic variables ($F(5,1225) = 82.02, p < 0.0001$). Using pairwise paired t -tests, and the Benjamini-Hochberg correction to control for the false discovery rate, 9 of the 15 topic-topic pairs are found to have a statistically significant difference (p -value < 0.0001) in mean expected replication rates. Results including significance groupings are reported in Figure 1B. The participants expect the highest replication rate in economics (average response 58%), and the lowest in Psychology and in Education (average response of 42% for both fields). The complete results are given in Supplementary Table 5.

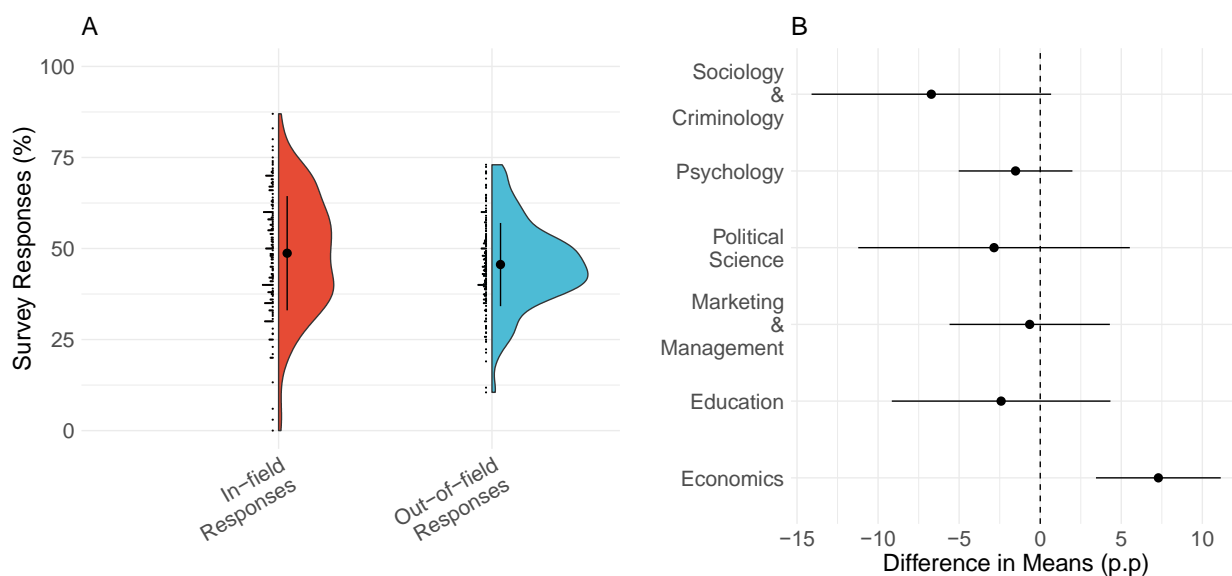
Figure 4-1. A. Expected replication rate for publications from different 2-year periods. B. Expected replication rate for publications from different fields. Points and error bars within the violin plots indicate the mean \pm one standard deviation. Letters in panel B indicate significance grouping: fields with the same grouping label do not have significantly different means. Groupings are omitted for panel A as all time periods have statistically significant or suggestive differences.



We observe that the forecasters' average responses for topics which match their own fields of interest are higher than their average responses for topics not belonging to their fields of interest (paired t -test, $t(197) = -3.35, p = 0.0010$). The average response for 'in-field' topics is 48.7%,

compared to 45.6% for ‘out-of-field’ topics (see Figure 2A). To identify the mechanisms behind this finding, we followed up with two additional analyses that were not pre-registered. To test if research fields are more optimistically assessed by participants with interest in this field, as compared to participants with no interest in this field, we performed unpaired t -tests, comparing ‘in-field’ responses with ‘out-of-field’ responses for each topic separately. The only statistically significant effect is found in economics. Participants interested in economics were more optimistic about the replication rates in economics than those not interested in economics (unpaired t -test, $t(199.85) = 3.74, p = 0.0002$). No evidence was found for such an effect within the other topics (see Figure 2B). Moreover, we investigated how the forecast for the overall replication rate depends on demographic characteristics. Suggestive evidence is observed for only one of the demographic variable included: participants who stated that they have been involved in a replication study before on average forecasted a lower overall replication rate ($t(217) = -2.75, p = 0.006$). Effect sizes and test statistics are given in Supplemental Table 6.

Figure 4-2. **A.** In-field vs. out-of-field responses. Participants predict a higher replication rate for their fields of interest, as compared to other fields. **B.** Difference of evaluation of a field by in-field and out-field participants (in percent points). Participants with interest in economics predict a higher replication rate for this field compared to participants with no interest in economics. For other fields, such an effect is not observed. Points and error bars indicate the mean \pm one standard deviation.



A key part of our experimental design for SCORE is the use of two alternative methodologies for eliciting and aggregating forecasts. Although there is an overlap of participants involved in survey and markets, the methodologies are independent and differ from each other in terms of elicitation and aggregation of information. To test if both methodologies yield similar results, we performed a correlation test between our two main aggregators: the smoothed market price and the forecasts generated by the survey-based peer assessment method. The Pearson correlation coefficient is 0.935 ($df = 10$, $p < 0.0001$), supporting that our findings are robust with respect to the elicitation and aggregation methodology.

4.5 Discussion

Forecasting research outcomes has been argued to benefit science (15,23,24). Hanson (15) suggested using prediction markets to forecast research outcomes to more efficiently reach a consensus on scientific questions and counteract inaccurate but popular beliefs. He also pointed out that funders could use prediction markets to incentivize research on questions they prioritize, without having to commit funding to specific research groups. Moreover, ex-ante predictions from prediction markets could help set priors for Bayesian statistical inference, prioritize research questions for hypotheses testing (23), and help to better capture how novel or surprising a result is (24).

The aim of our previous forecasting projects (8–11) was to test whether within the research communities there is information about the replicability of studies, and whether surveys and prediction markets can aggregate this information into accurate forecasts. The results of these previous studies were encouraging: the forecasted probabilities were informative with respect to the observed replication outcomes. For the SCORE project, we go beyond such a proof-of-principle.

We elicit information on a large set of research claims with only a small subset being evaluated through replication of reproduction. This approach illustrates how the information gained from the forecasting can be scaled up without necessarily scaling up cost-intensive replications.

The forecasts presented in this study focus on field-specific and time-specific replication rates, rather than the probability of replication for individual claims. Previous forecasting studies have shown that while forecasts for single claims are informative with respect to observed replication outcomes, they tend to be too optimistic. Explicit forecasts for overall replication rates might be more reliable than what can be inferred from forecasts on individual replications; this is a hypothesis we have pre-registered to test once the results from the SCORE replications are available.

Our results show that participants expect replication rates to increase over time, from 43% in 2009/10 to 55% in 2017/18. The reasons behind this expectation have not been elicited in our study, and are an interesting topic for future research. One plausible explanation might be that participants expect recent methodological changes in the social and behavioral sciences to have a positive impact on replication rates. This is also in line with an increased creation and use of study registries during this time period in the social and behavioral science (3). Further insights into longitudinal patterns in the reliability of published research could follow from replication projects on studies sampled across an extended period of time.

Similarly, the observed differences in topic-specific expected replication rates deserve further study. For fields that have been covered by replication studies in the past, the expected replication rates are likely anchored around past replication projects; the point estimate of the replication rate in the Replication Project: Psychology (22), for instance was lower than in the Experimental Economics Replication Project (8), although it should be noted that the inclusion

criteria, time periods, and sample sizes differed between these projects and thus it is not straightforward to compare these numbers. Differences in expected replication rates could further reflect that hypotheses and typical effect sizes differ between fields, different fields employ different methodologies and policies, and results from different fields might be subjected to different biases.

Because forecasts from previous replication reports were informative with respect to replication outcomes, the forecasts presented here might provide some guidance on how credible claims in different subjects are. This is particularly the case for fields for which no other information is available, i.e. fields with no past large scale replication project such as Education, Political Science and Marketing, Management and Related Areas. If our forecasts hold up, it will be interesting to investigate if specific factors (such as different methodologies and policies) can be identified that influence replication rates.

4.6 References

1. Fidler F, Wilcox J. Reproducibility of Scientific Results. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy. Winter 2018. Metaphysics Research Lab, Stanford University; 2018 [cited 2019 Nov 14]. Available from: <https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility/>
2. Schmidt S. Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Rev Gen Psychol.* 2009 Jun 1;13(2):90–100.
3. Christensen G, Miguel E. Transparency, Reproducibility, and the Credibility of Economics Research. *J Econ Lit.* 2018 Sep;56(3):920–80.
4. Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol Sci.* 2011 Nov 1;22(11):1359–66.
5. John LK, Loewenstein G, Prelec D. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychol Sci.* 2012 May 1;23(5):524–32.

6. Landy J, Jia M, Ding I, Viganola D, Tierney W, Dreber A, et al. Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychol Bull.* 2019 Oct 29 [cited 2020 Jan 20]; Available from: <http://repository.essex.ac.uk/25784/>
7. Coles NA, Tiokhin L, Scheel AM, Isager PM, Lakens D. The Costs and Benefits of Replication Studies. *PsyArXiv*; 2018 Jan [cited 2020 Jan 20]. Available from: <https://osf.io/c8akj>
8. Camerer CF, Dreber A, Forsell E, Ho T-H, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. *Science.* 2016 Mar 25;351(6280):1433–6.
9. Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nat Hum Behav.* 2018 Sep;2(9):637–44.
10. Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, et al. Using prediction markets to estimate the reproducibility of scientific research. *Proc Natl Acad Sci.* 2015 Dec 15;112(50):15343–7.
11. Forsell E, Viganola D, Pfeiffer T, Almenberg J, Wilson B, Chen Y, et al. Predicting replication outcomes in the Many Labs 2 study. *J Econ Psychol* 2018 Oct 25 [cited 2019 Mar 22]; Available from: <http://www.sciencedirect.com/science/article/pii/S0167487018303283>
12. Defense Sciences Office. Systematizing Confidence in Open Research and Evidence (SCORE). DARPADSO Broad Agency Announc HR001118S0047. 2018 Jun 12; Available from: <https://research-vp.tau.ac.il/sites/resauth.tau.ac.il/files/DARPA-SCORE-DSO-2018.pdf>
13. DARPA. DARPA SCORE Internal Documentation. 2019.
14. Nosek BA, Errington TM. What is replication? *MetaArXiv*; 2019 Sep [cited 2019 Oct 15]. Available from: <https://osf.io/u4g6t>
15. Hanson R. Could gambling save science? Encouraging an honest consensus. *Soc Epistemol.* 1995 Jan 1;9(1):3–33.
16. Hanson R. Combinatorial Information Market Design. *Inf Syst Front.* 2003 Jan 1;5(1):107–19.
17. Wolfers J, Zitzewitz E. Prediction Markets in Theory and Practice. National Bureau of Economic Research; 2006 Mar [cited 2019 Oct 16]. Report No.: 12083. Available from: <http://www.nber.org/papers/w12083>
18. Liu Y, Wang J, Chen Y. Surrogate Scoring Rules. *ACM Conf Econ Comput.* 2020 Jul; Available from: <http://arxiv.org/abs/1802.09158>
19. Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. *J Am Stat Assoc.* 2007 Mar 1;102(477):359–78.
20. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. *Nat Hum Behav.* 2018 Jan;2(1):6.

21. Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB, Alper S, et al. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Adv Methods Pract Psychol Sci*. 2018 Dec;1(4):443–90.
22. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015 Aug 28;349(6251):aac4716.
23. Pfeiffer T, Almenberg J. Prediction markets and their potential role in biomedical research--a review. *Biosystems*. 2010 Dec;102(2–3):71–6.
24. DellaVigna S, Pope D, Vivaldi E. Predict science to improve science. *Science*. 2019 Oct 25;366(6464):428–9.

4.7 Supplementary Material

Table 4-2. List of Journals and Discipline Clusters

Journal	Discipline Cluster
American Economic Journal: Applied Economics	Economics
American Economic Review	Economics
Econometrica	Economics
Experimental Economics	Economics
Journal of Finance	Economics
Journal of Financial Economics	Economics
Journal of Labor Economics	Economics
Journal of Political Economy	Economics
Quarterly Journal of Economics	Economics
Review of Financial Studies	Economics
American Educational Research Journal	Education
Computers and Education	Education
Contemporary Educational Psychology	Education
Educational Researcher	Education
Exceptional Children	Education
Journal of Educational Psychology	Education
Learning and Instruction	Education
Academy of Management Journal	Marketing, Management and Related Areas

Journal of Business Research	Marketing, Management and Related Areas
Journal of Consumer Research	Marketing, Management and Related Areas
Journal of Management	Marketing, Management and Related Areas
Journal of Marketing	Marketing, Management and Related Areas
Journal of Marketing Research	Marketing, Management and Related Areas
Journal of Organizational Behavior	Marketing, Management and Related Areas
Journal of Public Administration Research and Theory	Marketing, Management and Related Areas
Journal of the Academy of Marketing Science	Marketing, Management and Related Areas
Leadership Quarterly	Marketing, Management and Related Areas
Management Science	Marketing, Management and Related Areas
Organization Science	Marketing, Management and Related Areas
Organizational Behavior and Human Decision Processes	Marketing, Management and Related Areas
Public Administration Review	Marketing, Management and Related Areas
American Journal of Political Science	Political Science
American Political Science Review	Political Science
British Journal of Political Science	Political Science
Comparative Political Studies	Political Science
Journal of Conflict Resolution	Political Science
Journal of Experimental Political Science	Political Science
World Development	Political Science
World Politics	Political Science
Child Development	Psychology
Clinical Psychological Science	Psychology
Cognition	Psychology
European Journal of Personality	Psychology
Evolution and Human Behavior	Psychology
Health Psychology	Psychology
Journal of Applied Psychology	Psychology
Journal of Consulting and Clinical Psychology	Psychology
Journal of Environmental Psychology	Psychology
Journal of Experimental Psychology: General	Psychology
Journal of Experimental Social Psychology	Psychology

Journal of Personality and Social Psychology	Psychology
Psychological Medicine	Psychology
Psychological Science	Psychology
American Journal of Sociology	Sociology & Criminology
American Sociological Review	Sociology & Criminology
Criminology	Sociology & Criminology
Demography	Sociology & Criminology
European Sociological Review	Sociology & Criminology
Journal of Marriage and Family	Sociology & Criminology
Law and Human Behavior	Sociology & Criminology
Social Forces	Sociology & Criminology
Social Science and Medicine	Sociology & Criminology

Table 4-3. List of questions in the initial survey and market

- Q1) What will be the average replication rate in SCORE?
 Q2) What will be the average replication rate in economics?
 Q3) What will be the average replication rate in political sciences?
 Q4) What will be the average replication rate in psychology?
 Q5) What will be the average replication rate in education research?
 Q6) What will be the average replication rate in sociology and criminology?
 Q7) What will be the average replication rate in marketing, management and related areas?
 Q8) What will be the average replication rate for papers published in 2009/10?
 Q9) What will be the average replication rate for papers published in 2011/12?
 Q10) What will be the average replication rate for papers published in 2013/14?
 Q11) What will be the average replication rate for papers published in 2015/16?
 Q12) What will be the average replication rate for papers published in 2017/18?

Table 4-4. p-values for pairwise t-tests for time-specific responses (df = 225 for all tests)

	2009/10	2011/12	2013/2014	2015/2016
2011/2012	t = 2.741, p = 0.00662			
2013/2014	t = 10.6641, p < 0.00001	t = 6.3453, p < 0.00001		
2015/2016	t = 9.7299, p < 0.00001	t = 13.7561, p < 0.00001	t = 5.0903, p < 0.00001	
2017/2018	t = 14.9263,	t = 13.6037,	t = 14.6781,	t = 8.9061,

	p < 0.00001	p < 0.00001	p < 0.00001	p < 0.00001
--	-------------	-------------	-------------	-------------

Table 4-5. p-values for pairwise t-tests for topic-specific responses (df = 225 for all tests)

	Economics	Education	Marketing & Management	Political Science	Psychology
Education	t = -14.4446, p < 0.00001				
Marketing & Management	t = -12.866, p < 0.00001	t = 1.8142, p = 0.08190			
Political Science	t = -11.6888, p < 0.00001	t = 6.8161332, p < 0.00001	t = 5.0736, p < 0.00001		
Psychology	t = -16.202, p < 0.00001	t = 0.4283494, p = 0.66881	t = -1.2820, p = 0.21552	t = -7.3937, p < 0.00001	
Sociology & criminology	t = -11.5689, p < 0.00001	t = 4.3917, p = 0.00003	t = 1.8982, p = 0.07368	t = -3.9677, p = 0.00015	t = 3.2547, p = 0.00179

Table 4-6. Relation between forecast for the overall replication rate in score and demographic characteristics. The reference category for career stage is 'student'; category 'other' includes those who did not provide an answer as well as those who chose 'prefer not to answer' and 'other'. 'Academia' indicates those participants currently involved in academic activities either as a student or an employee. 'Prediction market' refers to participants who had been involved in a previous prediction markets. 'Replication' indicates that the participants have been involved in a replication study before. There is suggestive evidence that having being involved in a replication reduces the response for the overall SCORE replication rate forecast. An ANOVA indicates that the career stage variable has no statistically significant joined effect on the forecast ($F(4) = 0.396, p = 0.8115$).

term	estimate	SE	statistic	p value
Intercept	51.07218	5.569644	9.169739	3.75E-17
Career stage: early career	0.905805	2.441899	0.370943	0.711042
Career stage: mid career	2.388422	2.789189	0.856314	0.392769
Career stage: other	-3.54569	5.181558	-0.68429	0.494522
Career stage: senior career	0.093658	3.669664	0.025522	0.979662
Academia	-1.06542	5.251263	-0.20289	0.839413
Prediction market	2.101495	1.828185	1.149498	0.251616
Replication	-5.21526	1.894342	-2.75307	0.006404

5 Chapter 5: Creative destruction in science

This chapter contains the components of the paper “*Creative destruction in science*” that was published in 2020 in the journal “*Organizational Behavioral and Human Decision Processes*”. This paper describes a large project with many hypotheses tested and experiments conducted. I was involved only in the forecasting of the creative destruction replication results, and therefore only the parts relevant to the forecasting have been included in this chapter. My specific role was helping to inform the experimental and statistical design of the forecasting, undertaking the statistical analyses and report drafting and editing.

This project represents a clear extension of the forecasting found in chapters 2, 3 and 4. Here, we not only investigate the forecasting accuracy of predictions for direct replications, but also provide forecasts for multiple competing theories in the same space and test the extent to which forecasters can predict simple effects, moderator effects and interactions. In addition, we test if forecasters’ values regarding gender impact their ability to evaluate theories regarding gender.

The reference for this paper is: Tierney W, Hardy III JH, Ebersole CR, Leavitt K, Viganola D, Clemente EG, Gordon M, Dreber A, Johannesson M, Pfeiffer T, Uhlmann EL. Creative destruction in science. Organizational Behavior and Human Decision Processes. 2020 Nov 1;161:291-309.

To align with the formatting and referencing style of this thesis, there are some changes in formatting and referencing style of the published paper

5.1 Abstract

Drawing on the concept of a gale of creative destruction in a capitalistic economy, we argue that initiatives to assess the robustness of findings in the organizational literature should aim to simultaneously test competing ideas operating in the same theoretical space. In other words, replication efforts should seek not just to support or question the original findings, but also to replace them with revised, stronger theories with greater explanatory power.

Achieving this will typically require adding new measures, conditions, and subject populations to research designs, in order to carry out conceptual tests of multiple theories in addition to directly replicating the original findings. To illustrate the value of the creative destruction approach for theory pruning in organizational scholarship, we describe recent replication initiatives re-examining culture and work morality, working parents' reasoning about day care options, and gender discrimination in hiring decisions.

5.2 Forecasting Creative Destruction Replication Results

A complementary forecasting survey examined whether independent scientists were able to anticipate these replication results (see <https://osf.io/nz48k>, and Supplements, 7, 8, and 9 for the forecasting survey materials, pre-registered analysis plan, and detailed report). Prior work finds that scientists are able to accurately predict simple condition differences by merely reading the study abstract or examining the study materials (1–4). We tested, for the first time, whether scientists can likewise anticipate complex interactions between variables. In this politically charged context (5),

we further examined whether scientists' beliefs and values regarding gender moderate the accuracy of their predictions.

Consistent with past research, in our primary pre-registered hypothesis test, we found a positive association between the observed effect sizes and the individual predictions (beliefs) of the forecasters ($\beta = 0.027, p < 0.001$). In a pre-registered robustness test, aggregated predictions, computed as mean predicted effect size of each of the 24 effects replicated, were directionally positively associated with the observed effect sizes, although this zero-order correlation was no longer statistically significant, $r = 0.193, p = 0.366$. A notable discrepancy between forecasts about selection decisions by male evaluators and the actual study outcomes was also apparent. Forecasters expected that both male and female evaluators would prefer male job candidates (forecasted $d = 0.357$ for male evaluators; forecasted $d = 0.110$ for female evaluators, mean of the differences = 0.248, $p < 0.0001$). However, only the aggregate forecasts about selection decisions by female evaluators were in the same direction as the realized results (realized $d = -0.128$ for male evaluators; realized $d = 0.018$ for female evaluators). As a consequence, forecasters were less accurate at anticipating gender discrimination by male evaluators relative to female evaluators ($p < 0.0001$).

A non- preregistered follow up analysis revealed that 184 of 194 forecasters predicted that male evaluators would discriminate against female job candidates, directionally contrary to the replication results reported earlier (mean of the differences = 0.485, $p < 0.001$). Thus, although the expected positive association between forecasts and outcomes emerged for the moderator effects, for some simple effects the association is in the wrong direction (negative) and significant. Among forecasters, individual differences in beliefs about gender did not moderate accuracy (see Supplement 9). Further research should continue to examine whether scientists can predict the

results of complex experiments addressing socially sensitive topics, and what factors might facilitate (or impede) their accuracy.

5.3 Supplements for “Creative Destruction in Science”

5.3.1 Supplement 7: Pre-Registered Analysis Plan for the Forecasting Survey

Contributors to analysis plan: Domenico Viganola, Elena Giulia Clemente, Anna Dreber, Michael Gordon, Magnus Johannesson, Thomas Pfeiffer, Warren Tierney, Eric Luis Uhlmann.

Summary: In this survey, we will examine whether researchers can predict the results of a set of direct and conceptual replications of experimental research on gender and hiring decisions. We are targeting researchers with training in judgment and decision making/social psychology research to participate in the forecasting survey, with no exclusion based on seniority or any other demographic characteristic. Each participant (also referred to as forecaster in the rest of this pre-analysis plan) makes a total of $p = 24$ predictions. These will focus on the experimental effect sizes of the replications of hypotheses from Uhlmann & Cohen(6,7), as well as several novel effects derived from theories of gender discrimination. The predictions are subdivided into three groups:

- i. 2 predictions focusing on the simple effects (separately by evaluator gender)
- ii. predictions focusing on interaction effects (separately by evaluator gender)
- iii. 16 predictions focusing on moderator effects

In addition to making these predictions, the participants are asked to answer a set of questions aimed at eliciting their personal beliefs on gender-related topics as well as assessing their

demographics. Prior to data collection, the forecasting survey was piloted with a few colleagues to provide feedback on the clarity of the questions and design. The data for these pilot participants ($N = 8$) was not included in the final report as it occurred prior to the final preregistration of the methods and analyses.

In this forecasting study we use both the more conservative significance threshold of $p < 0.005$ (8) and the traditional threshold for statistical significance of $p < 0.05$. All the tests in this pre-analysis plan are two-sided tests.

5.3.1.1 Primary hypotheses

Hypothesis 1: There is a positive association between the predictions (beliefs) of the forecasters and the observed effect size

Individual-level regression to test whether forecasters' beliefs are significantly related to the realized effect sizes after controlling for individual fixed effects:

$$(1) \quad RES_p = \beta_0 + \beta_1 PES_{ip} + FE_i + \varepsilon_{ip}$$

where:

- RES_p is a continuous variable indicating the realized effect size of the hypothesis p object of the prediction;
- PES_{ip} is a continuous variable indicating the predicted effect size of the effect of hypothesis p of forecaster i ;
- FE_i is a set of individual fixed effects.

In equation (1) we plan to cluster standard errors at the individual level (number of clusters determined by the number of forecasters with $N = 24$ observations per cluster), since doing so allows us to take into account the fact that the predictions elicited from the same forecaster might be correlated.

Tests: t -test on coefficient β_1 in regression equation (1); t -test on coefficient β_0 in (1).

Robustness test of Hypothesis 1: we will estimate regression (1) separately for the three sets of predictions - predictions on simple effects, on interaction effects, and on moderator effects. Moreover, we will also carry out a robustness test where we estimate the Pearson correlation between the two vectors ($N = 24$ each) with the mean predicted effect size (PES_p) of each of the 24 effects replicated and the realized effect sizes RES_p .

Hypothesis 2

Can participants predict complex experimental results, such as interaction effects between conditions and individual differences moderators? To answer this question, first we compute the *accuracy* achieved in forecast p by each survey-taker i in terms of squared prediction error (Brier score), according to the formula:

$$BS_{ip} = (PES_{ip} - RES_p)^2$$

where RES_p and PES_{ip} should be interpreted as specified above. Then, we regress the variable BS_{ip} on 2 dummy variables identifying the forecasts regarding interactions ($INTES_{ip}$) and the forecasts regarding the effects of the moderators ($IDMES_{ip}$) and on the individual fixed effects FE_i , clustering the standard errors at the individual level in line with model (1):

$$(2) \quad BS_{ip} = \beta_0 + \beta_1 INTES_{ip} + \beta_2 IDMES_{ip} + FE_i + \varepsilon_{ip}$$

Tests: *t*-test on coefficient β_1 in regression equation (2); *t*-test on coefficient β_2 in (2); Wald test on coefficient β_1 being different from β_2 . Under the assumption that the forecasts on the interactions and on the moderators effects are more demanding, we expect both β_1 and β_2 to be positive.

5.3.1.2 Exploratory hypotheses

Introducing the ideological piece: how do scientists' political beliefs and convictions about gender relate to the accuracy of their forecasts? We exploit the individual accuracy measure introduced in hypothesis (2) and relate it to the forecasters' beliefs (sexist beliefs measure; beliefs about gender in the workplace; feminist media exposure measure; internal motivation to respond without sexism; external motivation to respond without sexism; political liberalism-conservatism on social issues) and to the forecasters' demographic characteristics (gender, academic seniority). The following tests are exploratory.

Individual-level regression to test whether forecasters' demographics and their convictions about gender relate to their accuracy in predicting the effect sizes. We plan to regress BS_{ip} on the following variables:

- Sexist beliefs measure (SBM_i)
- Feminist media exposure measure ($FMEM_i$)
- Beliefs about gender in workplace measure ($BGWM_i$)

- Internal motivation to respond without sexism ($IMSM_i$)
- External motivation to respond without sexism ($EMSM_i$)
- Political orientation on social issues measure (POL_i)
- Gender (G_i)
- Years from obtaining doctoral degree (SEN_i)

Please refer to the pre-registration document for the overall project (<https://osf.io/snbyg/>) and Supplements 2 and 4 for more details on these measures, most of which were also administered to the participants in the experiments whose results are being predicted.

Note that for these forecasts, we will again cluster the standard errors at the individual level to take into account potential correlations across forecasts made by the same forecaster:

$$(3) \quad BS_{ip} = \beta_0 + \sum^8 \beta_k IC_{ik} + \varepsilon_{ip} \quad \text{for } k = 1, \dots, 8$$

where $IC = \{SBM_i; FMEM_i; BGWM_i; IMSM_i; EMSM_i; POL_i; G_i; SEN_i\}$

Test: t -tests on coefficients β_1 to β_8 in regression equation (3).

As a robustness check for hypothesis 3, we will analyze the accuracy of predictions on simple effects, on interaction effects, and on moderators effects separately. Therefore, we will estimate the models in equation (3) on mutually exclusive subsets of all the predictions, namely:

- Predictions on gender discrimination patterns in hiring with $2 \times n$ observations, n being the total number of forecasters
- Predictions on interaction effects of experimental manipulations with $6 \times n$ observations
- Predictions on the moderators effect sizes with $16 \times n$ observations

Do predictions regarding gender discrimination in hiring by male evaluators differ from those regarding gender discrimination in hiring by female evaluators?

Are the predictions regarding the hiring evaluations made by women or men more accurate? We plan to answer this question by exploiting the fact that in the forecasting survey we ask exactly the same type of question for the two evaluator genders separately (e.g., ‘What do you predict will be the effect size for the influence of candidate gender on hiring evaluations among male participants?’ and ‘What do you predict will be the effect size for the influence of candidate gender on hiring evaluations among female participants?’). In order to test whether the predictions regarding discrimination by female and male evaluators differ significantly, we focus on the predictions of the simple effects as main test (1 test), and on the predictions of the interaction effects as secondary tests (3 tests). In the spirit of avoiding over-testing, we restrict the domain of these exploratory tests to the simple and the interaction effects, and to the differences in terms of predictions’ levels and predictions’ accuracy only.

Do the predictions about female and male evaluators differ significantly?

Test: paired t-test comparing the predictions regarding the simple effects about male evaluators and about female evaluators.

Test: paired t-test comparing the predictions regarding the interactions effects for male evaluators and for female evaluators, for a total of 3 different tests.

Do the predictions about female and male evaluators differ in terms of accuracy?

Test: paired t-test comparing the Brier score (BS_{ip} as defined for hypothesis 2) for predictions regarding the simple effects for male evaluators and for female evaluators.

Test: paired t-test comparing the Brier score for the predictions regarding the interactions effects for male evaluators and for female evaluators, for a total of 3 different tests.

5.3.1.3 Incentive scheme

The incentive scheme to participate in this study is composed of two parts: the first one is co- authorship on the study report and it is granted to all the forecasters; the second one is a monetary incentive granted to two forecasters who are randomly selected.

Co-authorship. Upon completion of the prediction survey in all its parts, the participants qualify to be listed as co-authors on the final manuscript reporting the results of this study, which will be submitted for publication in a scientific journal. The forecasters may join via a consortium credit (e.g., “Hiring Decisions Forecasting Collaboration”).

Monetary incentives. We will randomly select two of the participants and reward them with a bonus payout determined as a function of the accuracy of their forecasts. The bonus payoffs will be computed according to the following scoring rule:

$$\$200 - \frac{(Sq. Error \times 200)}{}$$

where *Sq. Error* is the average of the squared errors for all the 24 forecasts of the ‘Gender and Hiring Decisions Forecasting Study’ made by the forecasters.

5.4 Supplement 9: Detailed Report of the Forecasting Results

5.4.1 Methodological details

5.4.1.1 Materials.

We asked the respondents to the forecasting survey to each make a total of 24 predictions about effect sizes in terms of Cohen's d as well as the direction of the effect: two predictions focusing on simple effects of target gender (separately by evaluator gender), six predictions focusing on interaction effects (separately by evaluator gender), and 16 predictions focusing on moderator effects. Effect sizes were bounded between -3 and 3 . The forecasters were also asked to answer a set of questions capturing their personal beliefs on gender-related topics as well as assessing their demographics.

All the relevant study materials were fully disclosed to the forecasters, including detailed information about the sample sizes, sample characteristics, study design and materials (including links to complete study materials and pre-analysis plans), and links to the original articles targeted for replication.

5.4.1.2 Recruiting forecasters.

We targeted researchers with training in judgment and decision making/social psychology research to participate in the forecasting survey, with no exclusion based on seniority or any other demographic characteristic. We posted the link to a signup page for the forecasting survey on various academic websites, and online platforms and Facebook pages aimed at researchers in psychology, judgment and decision making and research methodology (e.g., Psych Map, Psych Methods Discussion Group, Judgment and Decision Making list). We also asked colleagues on

Twitter with many followers to post the link to the signup page. Once signing up, respondents received an individualized link to the forecasting survey. This link allowed them to start and continue with the survey at multiple occasions. Respondents also received at least two reminders to finish the survey.

Respondents were incentivized to participate in two ways: they were offered coauthorship on the study report via a consortium credit, and two randomly selected forecasters were rewarded with a bonus payment determined as a function of the accuracy of their forecasts using the following scoring rule: “\$200 - (Sq.Error 200)” where Sq.Error is the average of the squared errors for all the 24 forecasts of the ‘Gender and Hiring Decisions Forecasting Study’ made by the forecasters.

An initial group of 354 individuals signed up for the forecasting survey, out of which 194 completed the survey, while 111 started but did not complete the survey. 59.8% of the forecasters reported that they were men, 37.1% that they were women, and 1.5% chose ‘Other’ and 1.5% chose ‘Prefer not to tell.’ The average number of years after the PhD was 4.9 years (SD = 6.4). Note that the sample size and composition in an online survey of this kind is not under the control of the investigators. One has to accept whatever sample size and statistical power is achieved. Our final sample size was comparable to past academic forecasting surveys (e.g., Landy et al., (9)).

5.4.2 Results

Hypothesis tests. The planned analyses are outlined in our pre-analysis plan on <https://osf.io/nz48k/> and in Supplement 7. In the report below, we follow the pre-analysis plan unless otherwise specified.

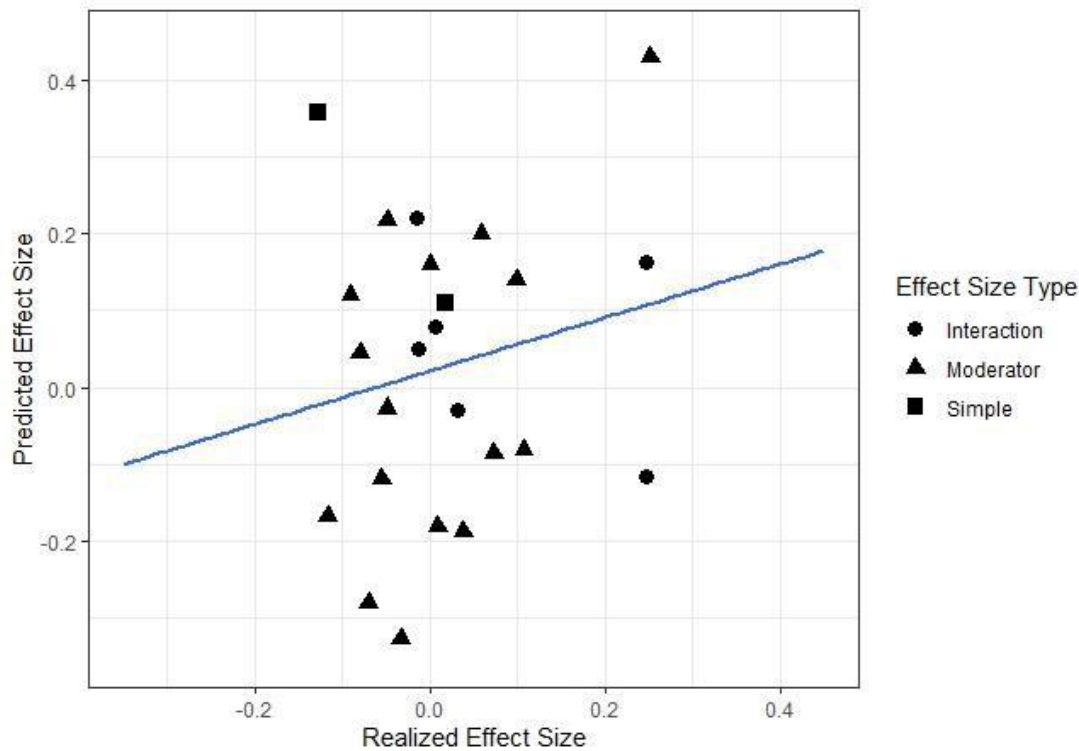
Our primary hypothesis 1 for the forecasting survey was that there would be a positive association between the predictions (beliefs) of the forecasters and the observed effect sizes. The individual-level regression and the t-test confirm that there is a positive and statistically significant association between the predictions of the forecasters and the observed effect sizes, with $\beta_1 = 0.027$ and $p < 0.0001$. See Table S9-1 for the individual-level regression estimates and Figure S9-1 for the correlation ($r = 0.193$, $p = 0.366$) between the average predicted effect sizes and the realized effect size.

Table 5-1: Correlation between forecasted and observed effect sizes.

<i>Dependent variable: Realized effect size</i>	
Forecasted effect size	0.027** (0.004)
Observations	4656
R ²	0.009

Note: * $p < 0.05$; ** $p < 0.005$. Standard errors clustered at individual level.

Figure 5-1 : Correlation between realized effect sizes and mean predicted effect sizes.



Our primary hypothesis 2 was that forecasters could predict complex experimental results, such as interaction effects between conditions and individual differences moderators. For this we compute the *accuracy* achieved in each forecast by each forecaster in terms of squared prediction error (Brier score). In the regression of the Brier score we find that both coefficients on the forecasts regarding interactions and the effects of the moderators are statistically significant but, contrary to expectations, negative, relative to predictions for simple effects. The coefficient on the variable identifying the forecasts regarding interaction effects is $\beta = -0.079$ with $p = 0.0002$ and that of the variable identifying the forecasts regarding the effects of the moderators is $\beta = -0.094$ with $p = 0.0036$. See Table S9-2.

Surprisingly, the results suggest that forecasters are able to predict experimental results and their accuracy is higher (lower Brier Score) for complex results such as interaction and moderator effects compared to simple effects. The Wald test cannot reject the null hypothesis that the two coefficients are equal ($p = 0.395$).

Table 5-2: Forecasts of interaction effects and moderators in terms of squared prediction error (Brier score).

<i>Dependent variable:</i>	
<i>Brier Score</i>	
Forecasts regarding interactions	-0.079** (0.017)
Forecasts regarding the effects of the moderators	-0.094** (0.016)
Observations	4656
R ²	0.008

Note: * $p < 0.05$; ** $p < 0.005$. Standard errors clustered at individual level.

5.4.2.1 Additional analyses.

We preregistered several ancillary exploratory hypotheses, all reported below in addition to one test that was not preregistered. As reported in the main text, we explore whether the forecasters' political beliefs and convictions about gender (sexist beliefs measure; beliefs about gender in the workplace; feminist media exposure measure; internal motivation to respond without sexism; external motivation to respond without sexism; political liberalism-conservatism on social issues; see supplements 2, 4, and 8 for more details on the measures) and the forecasters' demographic characteristics (gender where female is coded as 1 and the other three categories as 0, academic seniority measured by years since PhD) relate to the accuracy of their forecasts using the individual accuracy measure from hypothesis 2 (the Brier Score). Because there are so many of these individual- differences measures, we consider these analyses exploratory even though they were preregistered. See Table S9-3 for the summary statistics of the individual differences variables in the sample of forecasters.

Table 5-3: Summary statistics of measures in the exploratory hypotheses.

Variable	Mean	SD
Sexist beliefs measure	2.90	1.33
Feminist media exposure measure	5.05	1.13
Beliefs about gender in the workplace measure	5.52	1.06
Internal motivation to respond without sexism	5.785	1.11
External motivation to respond without sexism	3.10	1.67
Political orientation measure	2.57	1.20
Years since PhD	4.88	6.36

Further analyses indicate that none of the variables above are statistically significantly related to the accuracy of the forecast: sexist beliefs measure $\beta = -0.035$, $p = 0.275$, feminist media exposure $\beta = -0.015$, $p = 0.415$, beliefs about gender in the workplace measure $\beta = -0.014$, $p = 0.612$, internal motivation to respond without sexism measure $\beta = -0.002$, $p = 0.813$, external motivation to respond without sexism measure $\beta = -0.011$, $p = 0.182$, political orientation measure $\beta = 0.022$, $p = 0.095$, gender in the workplace measure $\beta = 0.028$, $p = 0.636$, and years since PhD measure $\beta = -0.006$, $p = 0.183$. See Table S9-4.

Table 5-4: Forecaster beliefs and demographics on squared prediction error (Brier Score).

	<i>Dependent variable:</i>
	<i>Brier Score</i>
Sexist beliefs measure	-0.035 (0.032)

Feminist media exposure measure	-0.015 (0.018)
Beliefs about gender in the workplace measure	-0.014 (0.028)
Internal motivation to respond without sexism	-0.002 (0.010)
External motivation to respond without sexism	-0.011 (0.008)
Political orientation measure	0.022 (0.013)
Female forecaster	0.028 (0.060)
Years since PhD	-0.006 (0.004)
Constant	0.412 (0.396)
<hr/>	
Observations	4656
R ²	0.013

*Note: *p < 0.05; **p < 0.005. Standard errors clustered at individual level.*

We also test whether predictions regarding gender discrimination in hiring by male evaluators differ from those regarding gender discrimination in hiring by female evaluators, in terms of levels and accuracy. This allows us to test whether the predictions about the hiring evaluations made by men or women are more accurate. In this analysis we only look at the predictions of the simple effect of candidate gender as the main test (one test), and on the predictions of the interaction effects as secondary tests (three tests). The results suggest that the predictions of simple effects and interactions effects are different for male and female evaluators (simple effect of candidate gender mean of the differences = 0.248 and $p < 0.0001$, affirmation-

threat mean of the differences = 0.112, $p = 0.002$, objectivity vs. neutral mindset mean of the differences = -0.085, $p = 0.007$, priming stereotypes vs. neutral concepts mean of the differences = 0.140, $p = 0.0003$). In terms of accuracy, respondents have less accurate predictions regarding the simple effect of candidate gender for male evaluators vs. female evaluators ($p < 0.0001$), and forecasters are again less accurate for male evaluators relative to female evaluators for two of the three interaction effects (affirmation-threat $p = 0.191$, objectivity vs. neutral mindset $p < 0.0001$, priming stereotypes vs. neutral concepts $p = 0.0005$).

5.4.2.2 Robustness tests.

We estimate hypothesis 1 separately for the three sets of predictions: predictions on simple effects, on interaction effects, and on moderator effects. For the predictions of simple effects there is a statistically significant negative correlation ($\beta = -0.150$ and $p = 0.0007$) with realized effect sizes, as well as for the interaction effects ($\beta = -0.034$, $p = 0.010$), while for the moderator effects the correlation remains positive and statistically significant ($\beta = 0.064$, $p < 0.0001$ respectively). See Table S9-5.

Table 5-5: Robustness test for hypothesis 1 for predictions on simple effects (1), interaction effects (2), and moderator effects (3) separately.

	<i>Dependent variable: Realized effect size</i>		
	(1)	(2)	(3)
Forecasted effect size	-0.150** (0.019)	-0.034** (0.011)	0.064** (0.004)
Observations	388	1164	3104
R ²	0.253	0.010	0.005

*Note: * $p < 0.05$; ** $p < 0.005$. Standard errors clustered at individual level.*

For hypothesis 1 we also pre-registered a robustness test where we estimate the Pearson correlation between the mean predicted effect size of each of the 24 effects replicated and the realized effect sizes. As noted in the main text, this correlation is positive (0.193) but not significant ($p = 0.366$).

For the exploratory hypothesis on whether forecasters' demographics and their convictions about gender relate to their accuracy in predicting the effect sizes we also estimate it separately for the three sets of predictions (predictions on simple effects, on interaction effects, and on moderator effects). We again find that none of the forecasters' characteristics is statistically significantly associated with their accuracy. See Table S9-6.

Table 5-6: Forecaster beliefs and demographics on squared prediction error (Brier Score) for predictions on simple effects, interaction effects and moderator effects separately.

	<i>Dependent variable: Brier Score</i>		
	(1)	(2)	(3)
Sexist beliefs measure	-0.026 (0.024)	-0.041 (0.030)	-0.033 (0.034)
Feminist media exposure measure	-0.028 (0.032)	-0.014 (0.022)	-0.014 (0.018)
Beliefs about gender in the workplace measure	0.017 (0.036)	-0.002 (0.028)	-0.022 (0.029)
Internal motivation to respond without sexism	-0.006 (0.015)	-0.006 (0.014)	0.000 (0.009)
External motivation to respond without sexism	-0.017 (0.016)	-0.009 (0.009)	-0.011 (0.008)
Political orientation measure	0.042 (0.041)	0.037 (0.020)	0.013 (0.010)
Female	0.160*	0.081	-0.007

	(0.077)	(0.064)	(0.063)
Years since PhD	-0.002	-0.004	-0.007
	(0.004)	(0.004)	(0.005)
Constant	0.281	0.316	0.464
	(0.268)	(0.366)	(0.429)
Observations	388	1164	3104
R ²	0.032	0.018	0.013

*Note: *p < 0.05; **p < 0.005. Standard errors clustered at individual level.*

We also carried out a regression that was not specified in the pre-analysis plan, where the focus is on whether forecasters' demographics and their convictions about gender relate to their accuracy in predicting the effect sizes on the simple effect of candidate gender among male evaluators only. Again we find no statistically associations with accuracy. In particular, forecasters' accuracy regarding gender discrimination by male evaluators was not associated with any of the following: forecasters' own sexist beliefs ($p = 0.380$), the feminist media exposure measure ($p = 0.939$), beliefs about gender in the workplace measure ($p = 0.897$), internal/external motivation to respond without sexism ($p = 0.478 / p = 0.735$), and political orientation ($p = 0.566$). See Table S9-7.

Table 5-7: Forecaster beliefs and demographics on squared prediction error (Brier Score) for main effect of candidate gender on male evaluators only.

<i>Dependent variable:</i>	
<i>Brier Score</i>	
Sexist beliefs measure	-0.023 (0.026)
Feminist media exposure measure	-0.002 (0.023)
Beliefs about gender in the	0.004

workplace measure	(0.030)
Internal motivation to respond without sexism	-0.015 (0.021)
External motivation to respond without sexism	-0.005 (0.016)
Political orientation measure	0.016 (0.027)
Female forecaster	0.200** (0.061)
Years since PhD	-0.005 (0.004)
Constant	0.387 (0.311)
<hr/>	
Observations	194
R ²	0.060

*Note: *p < 0.05; **p < 0.005. Standard errors clustered at individual level.*

5.5 References

1. Camerer CF, Dreber A, Forsell E, Ho T-H, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. *Science*. 2016 Mar 25;351(6280):1433–6.
2. DellaVigna S, Pope D, Vivaldi E. Predict science to improve science. *Science*. 2019 Oct 25;366(6464):428–9.
3. Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, et al. Using prediction markets to estimate the reproducibility of scientific research. *Proc Natl Acad Sci*. 2015 Dec 15;112(50):15343–7.
4. Forsell E, Viganola D, Pfeiffer T, Almenberg J, Wilson B, Chen Y, et al. Predicting replication outcomes in the Many Labs 2 study. *J Econ Psychol*. 2019 Dec 1;75:102117.
5. Tetlock PE. *Expert Political Judgment: How Good Is It? How Can We Know?* Expert Political Judgment. Princeton University Press; 2009 [cited 2021 Oct 1]. Available from: <https://www.degruyter.com/document/doi/10.1515/9781400830312/html>
6. Uhlmann EL, Cohen GL. Constructed Criteria: Redefining Merit to Justify Discrimination. *Psychol Sci*. 2005 Jun 1;16(6):474–80.
7. Uhlmann EL, Cohen GL. “I think it, therefore it’s true”: Effects of self-perceived objectivity on hiring discrimination. *Organ Behav Hum Decis Process*. 2007 Nov 1;104(2):207–23.
8. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. *Nat Hum Behav*. 2018 Jan;2(1):6.
9. Landy J, Jia M, Ding I, Viganola D, Tierney W, Dreber A, et al. Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychol Bull*. 2019 Oct 29 [cited 2020 Jan 20]; Available from: <http://repository.essex.ac.uk/25784/>

6 Chapter 6: A creative destruction approach to replication: Implicit work and sex morality across cultures

This chapter contains part of the paper “*A creative destruction approach to replication: Implicit work and sex morality across cultures*” that was published in 2021 in the journal ‘*Journal of Experimental Social Psychology*’. As in chapter 5, this paper also describes a large project with a large team. My role was limited to the forecasting component. Therefore, only the sections that I contributed to are included in this thesis. I contributed to the experimental and statistical design of the forecasting, undertaking the statistical analyses, and report drafting and editing. This paper seeks to not only forecast direct replications (such as chapter 2 and the SCORE project -as described in appendix 1) but also investigates whether forecasters can predict the effect that cultural context will have on replication outcomes including effect sizes.

The reference for this paper is: Tierney W, Hardy III J, Ebersole CR, Viganola D, Clemente EG, Gordon M, Hoogeveen S, Haaf J, Dreber A, Johannesson M, Pfeiffer T. A creative destruction approach to replication: Implicit work and sex morality across cultures. Journal of Experimental Social Psychology. 2021 Mar 1;93:104060.

To align with the formatting and referencing style of this thesis, there are some changes in formatting and referencing style of the published paper

A creative destruction approach to replication: Implicit work and sex morality across cultures

6.1 Abstract

How can we maximize what is learned from a replication study? In the creative destruction approach to replication, the original hypothesis is compared not only to the null hypothesis, but also to predictions derived from multiple alternative theoretical accounts of the phenomenon. To this end, new populations and measures are included in the design in addition to the original ones, to help determine which theory best accounts for the results across multiple key outcomes and contexts. The present pre-registered empirical project compared the Implicit Puritanism account of intuitive work and sex morality to theories positing regional, religious, and social class differences; explicit rather than implicit cultural differences in values; self-expression vs. survival values as a key cultural fault line; the general moralization of work; and false positive effects. Contradicting Implicit Puritanism's core theoretical claim of a distinct American work morality, a number of targeted findings replicated across multiple comparison cultures, whereas several failed to replicate in all samples and were identified as likely false positives. No support emerged for theories predicting regional variability and specific individual-differences moderators (religious affiliation, religiosity, and education level). Overall, the results provide evidence that work is intuitively moralized across cultures.

6.2 Forecasting Survey

Given the findings from both Studies 1 and 2 are quite contrary to the original theorizing (1–3), an interesting question is whether the replication results are predictable by psychologists and other scholars. In a forecasting survey accompanying the present project, independent scientists were provided with descriptions of the competing theories and asked to try to predict the replication effect

sizes associated with each targeted effect. Two hundred and twenty-one colleagues made predictions about the target age and needless work effect, needless work main effect (works vs. retires) in the same “postal worker” scenario, tacit inference effect, intuitive work morality effect, and salvation prime effect, across each online sample for which data was collected (MTurk: USA and India; PureProfile: New England U.S. states, non-New-England U.S. states, Australia, and United Kingdom). For each targeted effect, we also asked forecasters to predict the aggregated effect size across samples for four key theoretical moderators: participant religious affiliation (Protestant or not), religiosity (DUREL score), Protestant work ethic endorsement, and education level.

Prior investigations demonstrate that scientists can anticipate simple condition differences based on mere examination of study abstracts or materials (4–7). We examined, for the first time, whether they can likewise accurately predict empirical outcomes when the same research paradigms are repeated in multiple cultural contexts. See <https://osf.io/7uhcg/> and Supplements, 4, 5, and 6 for the forecasting survey pre-registered analysis plan, survey materials, and a detailed report of the results. Summarizing briefly, in our primary hypothesis test, we found a statistically significant positive overall association between realized and predicted effect sizes, $\beta = 0.157$, $p = 0.0005$. The Pearson correlation between the mean predicted effect size of each of the 48 effects replicated and the observed effect sizes was likewise significant, $r = 0.704$, $p < 0.0001$. Thus, even when the pattern of results being predicted is quite complex, the accuracy of scientific forecasters remains a robust phenomenon (8,9).

At the same time, comparing the absolute differences between the forecasted and realized effect sizes (Cohen’s d) for each original effect underscores that this accuracy was less than perfect. Specifically, forecasted effect sizes averaged across populations were significantly different from the realized effect sizes, aggregated for each key effect via a random effect meta-analysis, for two of the five key effects at the $p < .005$ level (10) and for a third effect at the traditional $p < .05$ level. For the needless work main effect (works vs. retires), mean forecasts = 0.3233, and meta-analyzed realized effect size = 0.6524, with the difference between the two statistically significant, $p < 0.0001$, such

that participants underestimated the replication effect size. Forecasters likewise believed the tacit inferences effect would be smaller than it turned out to be, mean forecasts = 0.3114, meta-analyzed effect size = 0.5053, $p = 0.0055$. In contrast, for the target age moderating needless work effect, participants systematically overestimated the effect size, mean forecasts = 0.2461, meta-analyzed realized effect size = 0.032, $p < 0.0001$, believing the effect would replicate when in fact it did not. Forecasters expected a small but significant overall salvation prime effect, mean forecasts = 0.0972, which did not emerge, meta-analyzed effect size = 0.0104, but the difference between forecasted and realized effect sizes was not statistically significant, $p = 0.9181$. Finally, for the intuitive work morality effect, mean forecasts = 0.2520, were closely aligned with the meta-analyzed realized effect size = 0.2568, with no significant difference between them, $p = 0.954$.

Overall, forecasters did quite well in anticipating the replication outcomes, although they were less accurate in predicting absolute effect sizes than their direction and relative ordering. Based on their pattern of forecasted results, these independent scientists appear to have endorsed the general moralization of work theoretical perspective, in that they forecasted all the original effects would emerge and further would do so across cultures (see Tables S6-3 and S6-7 in Supplement 6). For the most part this facilitated successful forecasts, the general moralization of work being the most empirically supported theory in this replication initiative. The major exceptions are of course the salvation prime effect and target age and needless work effect, which failed to replicate as anticipated by the false positives account. Further research should continue to examine the extent to which scientists are able to anticipate cross-cultural replication results, ideally using a larger number of cultural populations than the relatively small set sampled here, as well as effects that exhibit greater heterogeneity across societies.

6.3 Supplementary Materials

6.3.1 CULTURE AND WORK REPLICATION PROJECT: PRE-ANALYSIS PLAN FOR THE FORECASTING SURVEY

Contributors to analysis plan: Domenico Viganola, Elena Giulia Clemente, Anna Dreber, Michael Gordon, Magnus Johannesson, Thomas Pfeiffer, Warren Tierney, Jay Hardy, Charlie Ebersole, Eric Luis Uhlmann.

Summary: In this survey we will examine whether researchers can predict the extent to which experimental findings regarding work morality replicate in data collections in different cultures and populations around the world. Of particular interest is the tendency to morally praise individuals for working in the absence of material need to work (the “needless work” effect), as well as linking work to other forms of traditional morality and divine salvation (1,2). The data for the replications are collected in the United States (differentiating the New England states from the rest of the country), United Kingdom, Australia, and India.

We are targeting researchers with training in judgment and decision making/social psychology research to participate in the forecasting survey, with no exclusion based on seniority or any other demographic characteristic.

Each participant (also referred to as forecaster in the rest of this pre-analysis plan) makes a total of $p = 48$ predictions. These will focus on five different work morality effects:

1. Needless work effect - 6 predictions regarding effect sizes in different populations and 4 predictions regarding moderator effects
2. Target age effect - 6 predictions regarding effect sizes in different

- populations and 4 predictions regarding moderator effects
3. Intuitive work morality effect - 6 predictions regarding effect sizes in different populations and 4 predictions regarding moderator effects
 4. Tacit inferences effect - 6 predictions regarding effect sizes in different populations and 4 predictions regarding moderator effects
 5. Salvation primes and work behavior - 4 predictions regarding effect sizes in different populations and 4 predictions regarding moderator effects

The data for these direct and conceptual replications are collected in the USA as a whole (MTurk sample), USA New England states (PureProfile sample), USA non-New England states (PureProfile sample), UK (PureProfile sample), Australia (PureProfile sample), and India (MTurk sample) for effects #1-4. For the fifth effect, no MTurk data was collected, hence the predictions are for USA New England States, USA non-New-England states, Australia, and UK, all sampled via the professional survey firm PureProfile. In addition to making these predictions, the participants are asked to answer a set of demographic questions.

Prior to data collection, the forecasting survey was piloted with a few colleagues to provide feedback on the clarity of the questions and design. The data for these pilot participants was not included in the final report as it occurred prior to the final preregistration of the methods and analyses, and we also revised the survey in light of the pilot feedback.

In this forecasting study we use both the more conservative significance threshold of $p < 0.005$ proposed by Benjamin et al. (2018) and the traditional threshold for statistical significance of $p < 0.05$. All the tests in this pre-analysis plan are two-sided tests.

6.3.1.1 Primary Hypotheses

Primary Hypothesis 1: There is a positive association between the predictions (beliefs) of the forecasters and the observed effect sizes

Individual-level regression to test whether forecasters' beliefs are significantly related to the realized effect sizes:

$$(1) \quad RES_{hc} = \beta_0 + \beta_1 PES_{ihc} + \varepsilon_{ihc}$$

where:

RES_{hc} is a continuous variable indicating the realized effect size of the hypothesis h object of the prediction in population c ; PES_{ihc} is a continuous variable indicating the predicted effect size of the effect of hypothesis h in population c by forecaster i ;

In equation (1) we plan to cluster standard errors at the individual level (number of clusters determined by the number of forecasters with $N = 48$ observations per cluster), since doing so allows us to take into account the fact that the predictions elicited from the same forecaster might be correlated.

Tests: t -test on coefficient β_1 in regression equation (1).

Robustness test of Hypothesis 1: we will estimate regression (1) separately for the two sets of predictions - predictions regarding simple effects and regarding moderator effects. Moreover, we will also carry out a robustness test where we estimate the Pearson correlation between the two vectors ($N = 48$ each) with the mean predicted effect size (PES_{hc}) of each of the 48 effects

replicated and the realized effect sizes RES_{hc} . Finally, we will estimate the Pearson correlation separately for the predictions regarding simple effects and the predictions regarding the moderator effects.

Primary Hypothesis 2: Forecasts regarding simple effect sizes are more accurate than forecasts regarding moderator effect sizes

Can participants predict complex experimental results, such as those associated with each candidate moderator, with the same accuracy achieved in predictions of simple effect sizes? To answer this question, first we compute the *accuracy* achieved in forecast hc by each survey-taker i in terms of squared prediction error (Brier score), according to the formula:

$$BS_{ihc} = (PES_{ihc} - RES_{hc})^2$$

where RES_{hc} and PES_{ihc} should be interpreted as specified above. Then, we regress the variable BS_{ihc} on a dummy variable identifying the forecasts regarding moderators (MES_{ihc}) and on the individual fixed effects FE_i , clustering the standard errors at the individual level in line with model (1).

$$(2) \quad BS_{ihc} = \beta_0 + \beta_1 MES_{ihc} + FE_i + \varepsilon_{ihc}$$

Tests: t -test on coefficient β_1 in regression equation (2). Under the assumption that the forecasts regarding the moderators effects are more demanding, we expect β_1 to be positive.

6.3.1.2 Secondary Hypothesis

Secondary hypothesis: Forecasted effect sizes are not significantly different from the realized effect sizes.

Hypothesis 1 tests the correlation between forecasts and realized effect sizes, but is not informative about the difference between the realized effects and their forecasted counterparts. To investigate whether the forecasted effect sizes are significantly different from the realized ones, we plan to apply the following procedure. First, for each of the 5 key effects we estimate the meta-analytic mean effect size PES^m_h , h ranging between 1 and 5, by pooling the effect sizes across the different cultures and populations (namely, across 6 populations for key effects 1 to 4 and across 4 populations for effect 5, as specified above) in a random effects meta-analysis. Then, we estimate the average at the individual level of the effect size of each key effect across the different populations for each participant (PES_{ih}). Finally, for each of the five key effects we implement a z-test comparing the meta-analyzed effect size PES^m_h to the mean of PES_{ih} .

6.3.1.3 Exploratory Hypotheses

Do participants predict experimental results across different populations with different degrees of accuracy? To answer this question we plan to estimate equation (3):

$$(3) \quad BS_{ihc} = \beta_0 + \beta_1 USNE_c + \beta_2 USNNE_c + \beta_3 US_c + \beta_4 AUS_c + \beta_5 IN_c + FE_i + \varepsilon_{ihc}$$

where BS_{ihc} and FE_i should be interpreted as above and $USNE_c$, $USNNE_c$, US_c , AUS_c and IN_c are dummy variables identifying forecasts on New England states in US (data collected via PureProfile), non-New England states in US (PureProfile), US (MTurk), Australia (PureProfile),

and India (MTurk) respectively (United Kingdom being the baseline population). In line with previous regressions, in equation (3) the standard errors are clustered at the individual level.

Tests: separate t -test on coefficients β_1 to β_5 in regression equation (3); Wald test on coefficients β_i being different from β_j for $i, j \in (1,2,3,4,5)$.

As a robustness check for the exploratory hypothesis we will analyze the accuracy of predictions on simple effects and on moderators effects separately. Therefore, we will estimate the model in equation (3) on two mutually exclusive subsets of all the predictions, namely:

- Predictions regarding the five key work morality effect sizes
- Predictions regarding the four moderator effects

Are the forecasters' years of academic experience related to higher accuracy? To answer this question, we plan to regress BS_{ihc} on the variable SEN_i which represents the year from when the PhD was awarded (this variable takes value zero if a PhD title is not awarded yet). We will again cluster the standard errors at the individual level to take into account potential correlations across forecasts made by the same forecaster.

$$(4) \quad BS_{ihc} = \beta_0 + \beta_1 SEN_i + \varepsilon_{ihc}$$

Test: t -tests on coefficient β_1 in regression equation (4).

As a robustness check for hypothesis 3, we will analyze the accuracy of predictions on simple effects and on moderators effects separately. We will also use a different proxy of seniority, namely, academic job rank.

6.3.1.4 Incentives scheme

The incentive scheme to participate in this study is composed of two parts: the first one is co- authorship on the study report and it is granted to all the forecasters; the second one is a monetary incentive granted to two forecasters who are randomly selected.

Co-authorship. Upon completion of the prediction survey in all its parts, the participants qualify to be listed as co-authors on the final manuscript reporting the results of this study, which will be submitted for publication in a scientific journal. The forecasters may join via a consortium credit (e.g., “Work and Culture Forecasting Collaboration”).

Monetary incentives. We will randomly select two of the participants and reward them with a bonus payout determined as a function of the accuracy of their forecasts. The bonus payoffs will be computed according to the following scoring rule:

$$\$200 - \frac{(Sq. Error \times 200)}{\quad}$$

where *Sq. Error* is the average of the squared errors for all the 48 forecasts of the ‘Work and Culture Forecasting Study’ made by the forecasters.

6.3.2 Supplement 6: Detailed Report of the Forecasting Results

Methodological details

6.3.2.1 Materials.

Respondents (forecasters) to the forecasting survey were asked to each make a total of 48 predictions regarding five different work morality effects (‘key effects’) in terms of effect sizes (Cohen’s d) and the direction of the effect. Twenty-eight predictions were regarding effect sizes in

different populations and 20 predictions regarding moderator effects. Effect sizes were bounded between -3 and 3. Forecasters were also asked to answer a set of demographic questions including their PhD year and job rank. Forecasters could access all the relevant study materials. These included detailed information about the sample sizes, sample characteristics, study design and materials, including links to the original articles and the complete study materials and pre-analysis plans for the replication.

6.3.2.2 Recruiting forecasters.

As in our other forecasting projects, we targeted researchers with training in judgment and decision making and/or social psychology research to participate in the forecasting survey. We excluded no respondents based on e.g. seniority or any other demographic characteristic. The link to a signup page for the forecasting project was posted on various academic websites, platforms, and Facebook pages aimed at researchers in psychology, judgment and decision making, and research methodology (e.g. Psych Map, Psych Methods Discussion Group, Judgment and Decision Making list). Colleagues with large followings on Twitter were also asked to post the link to the signup page. After having signed up, respondents received an individualized link to the forecasting survey, which allowed them to complete the survey in multiple sittings if they wished. Respondents received at least two reminders to finish the survey.

We incentivized participation in two ways. First, forecasters were offered coauthorship on the manuscript through a consortium credit ('Culture & Work Forecasting Collaboration'). Second, two forecasters were randomly selected and monetarily rewarded based on the accuracy of their forecasts using the following scoring rule:

$$\$200 - \frac{(Sq. Error \times 200)}{\quad}$$

where $Sq. Error$ is the average of the squared errors for all the 48 forecasts of the ‘Culture & Work Forecasting Study’ made by the forecasters.

Initially, 429 individuals signed up for the forecasting survey, out of which 222 completed the survey. One hundred and fifty of the individuals who had initially signed up started but did not ultimately complete the survey, and 57 signed up but never started their forecasts. One forecaster was removed from the sample for a technical issue that rendered her/his data unusable. Therefore, the final set of forecasters includes 221 respondents. This final sample size is comparable to past academic forecasting surveys (8,9). In terms of gender, 38.9% of the forecasters reported that they were women, 59.7% that they were men, 0.005% chose ‘other’ and 0.01% chose ‘prefer not to tell.’ Forecasters reported 48 countries of birth and 38 countries of residence. Out of 221 forecasters, 72% of them were born in either Europe (100 forecasters) or North America (58 forecasters), and 80% of them currently reside in Europe (96 forecasters) or North America (80 forecasters). The most represented countries of birth were the United States with 50 forecasters, Germany with 20 and the United Kingdom with 10, while the most represented countries of residence were the United States with 72 forecasters, the United Kingdom with 20, and the Netherlands with 13. The average number of years since PhD was four ($SD = 5.6$). Given the nature of the recruitment method (social media), sample size (and thus statistical power) as well as the sample composition were not under our full control. We simply tried to recruit as many forecasters as we could within the pre-registered time frame for data collection.

6.3.2.3 Results

The planned analysis is reported in our pre-analysis plan on <https://osf.io/7uhcg/> and in Supplement 4. We follow the pre-analysis plan unless otherwise specified.

Our primary hypothesis 1 was that there would be a positive association between the predictions (beliefs) of the forecasters and the observed effect sizes. As expected, the individual-level regression and t-test show a positive association between the predictions of the forecasters and the observed replication effect sizes, $\beta_1 = 0.157$, $p = 0.0008$. See Table S6-1 for the individual-level regression estimates.

Table 6-1. Association between forecasted and observed effect sizes.

<i>Dependent variable: Realized effect size</i>	
Forecasted effect size	0.157** (0.045)
Observations	10608
R^2	0.037

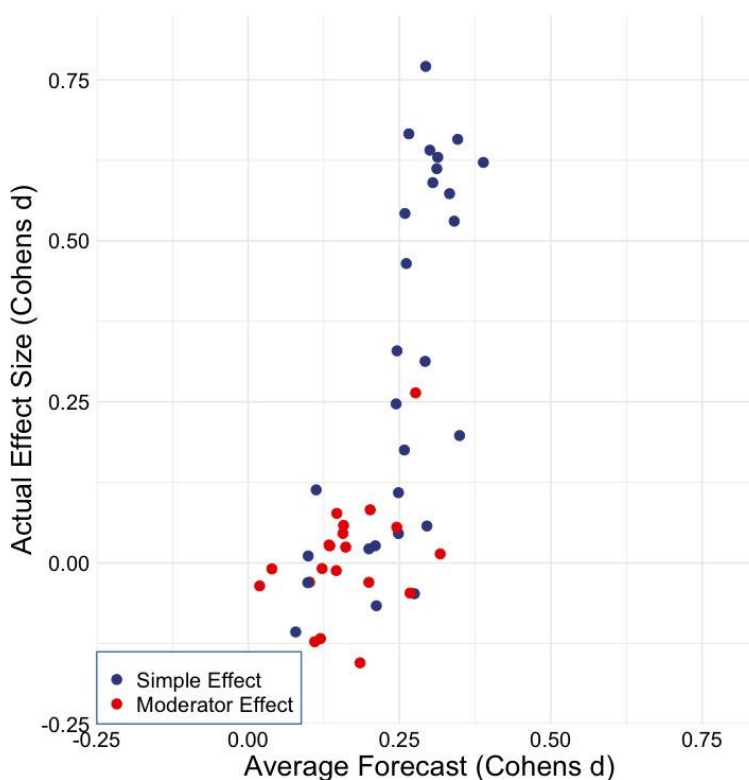
*Note: * $p < 0.05$; ** $p < 0.005$. Standard errors clustered at individual level.*

As a robustness test, we estimate hypothesis 1 separately for the two sets of predictions (simple effects and moderator effects). Focusing on simple effects only, there is a positive association ($\beta = 0.164$, $p = 0.002$). For the moderators alone, the association between predictions and effect sizes is significant using the traditional p-value cutoff of .05, but not the stricter .005 significance threshold proposed by Benjamin et al. (10) for which it represents suggestive evidence ($\beta = 0.019$, $p = 0.04$).

In another robustness test, we estimate the Pearson correlation between the mean predicted effect size of each of the 48 effects replicated and the observed effect sizes. Figure S6-1 displays the correlation ($r = 0.704$, $p < 0.0001$) between the average predicted effect sizes and the observed effect size. We also estimate the Pearson correlation separately for the predictions regarding simple

effects ($r = 0.688, p < 0.0001$) and the predictions regarding the moderator effects ($r = 0.375, p = 0.104$). The correlations for all effects combined and the simple effects separately are large and significant, but the correlation for moderator effects separately is not found to be statistically significant. This suggests that forecasters are for the most part able to anticipate the realized effect sizes, but their accuracy is not perfect. Further research is needed to establish whether or not forecasters are able to accurately predict the moderators of replication effect sizes.

Figure 6-1. Actual effect size vs average forecast (Cohen's d). Correlation between forecasted and actual effects for both simple and moderator effects (differentiated by the colors blue and red).



Our primary hypothesis 2 was that forecasters would be able to predict simple effect sizes more accurately than moderator effect sizes. For this we compute the *accuracy* achieved in each prediction by each forecaster in terms of squared prediction error (Brier score). In the regression of the Brier score we find no evidence for a relationship between effect type and accuracy (see Table

S6-2). The coefficient for the variable identifying the forecasts regarding moderator effects is $\beta = 0.008$, $p = 0.5221$. Thus, we cannot conclude that forecasters are significantly better at predicting simple effects than moderator effects.

Table 6-2: Forecasts of moderator effects relative to simple effects in terms of squared prediction error (Brier score).

<i>Dependent variable: Brier Score</i>	
Forecasts for moderator effects	0.008 (0.013)
Observations R^2	10608 0.0001

*Note: * $p < 0.05$; ** $p < 0.005$. Standard errors clustered at the individual level.*

While our primary hypothesis 1 tests the correlation between forecasts and realized effect sizes, it does not take into account the absolute difference between them. Our secondary hypothesis was that forecasted effect sizes would not be statistically significantly different from the realized effect sizes. We compare the meta-analyzed effect size for each of the five key effects (by pooling the effect sizes across the different cultures and populations) to the mean at the individual level of the effect size of each key effect (across the different populations for each participant). We then implement a z-test comparing whether these are statistically significantly different. The results are summarized in Table S6-3.

Table 6-3: Summary of the differences between meta-analyzed effect sizes and forecasts (standard errors in parenthesis)

Effect	Meta-analyzed effect	Mean of the forecasts	Difference	P-value
---------------	-----------------------------	------------------------------	-------------------	----------------

Needless work main effect (works vs. retires)	0.652 (0.031)	0.323 (0.014)	0.329	<0.0001
Target age and needless work effect	0.032 (0.041)	0.246 (0.017)	0.214	<0.0001
Intuitive work morality effect	0.257 (0.082)	0.252 (0.015)	0.005	0.954
Tacit inferences effect	0.505 (0.068)	0.311 (0.017)	0.1939	0.0055
Salvation prime and work behavior	0.010 (0.844)	0.097 (0.012)	0.087	0.9181

For two key effect sizes out of five, the main effect of needless work (works vs. retires) and target age and needless work effect, the mean of the forecasts and the meta analyzed effects are statistically significantly different from each other at the .005 level (10), with forecasts underestimating the former effect and overestimating the latter one. For the tacit inferences effect forecasters significantly underestimate the effect using the traditional .05 significance criterion, but not the more conservative .005 criterion proposed by Benjamin et al. (10). The z-tests for salvation prime and work behavior and the intuitive work morality effect fail to reject the null hypothesis that the means of the forecasts and the meta-analyzed effects are not statistically different.

We prespecified further analyses we regard as exploratory given the number of statistical tests involved and lack of strong theoretical predictions. First, we test whether forecasters can predict experimental results across different populations with different degrees of accuracy. In a regression we have binary variables for the New England states in the USA (data collected via

PureProfile), non-New England states in the USA (PureProfile), USA (MTurk), Australia (PureProfile), and India (MTurk) respectively, with United Kingdom being the baseline population. We do separate t-tests on the coefficients for these binary variables (β_1 to β_5) and a set of Wald tests on whether these coefficients are pairwise statistically significantly different. As Table S6-4 shows, we find that accuracy varies statistically significantly across some locations compared to the United Kingdom baseline population ($\beta_1 = -0.008, p = 0.351$; $\beta_2 = -0.023, p = 0.188$; $\beta_3 = 0.083, p < 0.0001$; $\beta_4 = 0.016, p = 0.0003$; $\beta_5 = 0.042, p < 0.0001$). The set of pairwise Wald tests summarized in Table S6-5 indicate that we cannot reject the null hypothesis that the coefficients are the same for two pairs of populations among these pairwise tests: New England states/non-New England states in the US and USA/India.

Table 6-4: Regression estimates of accuracy on country indicators.

<i>Dependent variable: Brier Score</i>	
USA – New England States (PureProfile)	-0.008 (0.008)
USA – Non-New England States (PureProfile)	-0.023 (0.017)
USA (MTurk)	0.083** (0.021)
Australia (PureProfile)	0.016** (0.004)
India (MTurk)	0.042* (0.009)
Observations	6188
R ²	0.009

*Note: *p < 0.05; **p < 0.005. Standard errors clustered at individual level.*

Table 6-5: P-values resulting from pairwise Wald tests on country coefficients shown in Table S6-4 being different from each other.

	USA – New England States (PureProfile)	USA – Non- New England States (PureProfile)	USA (MTurk)	Australia (PureProfile)	India (MTurk)
USA – New England States (PureProfile)	-	0.1704	<0.0001	0.0016	<0.0001
USA – Non- New England States (PureProfile)	-	-	0.0007	0.0100	<0.0001
USA (MTurk)	-	-	-	0.0026	0.1251
Australia (PureProfile)	-	-	-	-	0.0075
India (MTurk)	-	-	-	-	-

Finally, in an exploratory vein, we test whether the forecasters' years of academic experience (i.e., years since PhD) are related to higher accuracy. The results from the regression and the t-test on the seniority coefficient indicate that years since PhD is not statistically significant correlated with accuracy ($\beta_1 = 0.00024$, $p = 0.96$). As a robustness check for this exploratory hypothesis, we analyze the accuracy of predictions on simple effects and on moderator effects separately. We find a similar result for simple effects ($\beta_1 = 0.001$, $p = 0.724$) and moderator effects ($\beta_1 = -0.001$, $p = 0.843$). Also, as a robustness check, we use academic job rank as a different proxy

of seniority. We find that none of the academic ranks has a statistically significant correlation with accuracy relative to the reference group, i.e. those who selected “other” as job rank (see Table S6-6).

Table 6-6: Regression estimating the effects of academic seniority on forecasting accuracy.

	<i>Dependent variable: Brier Score</i>		
	(1) Full Sample	(2) Simple effects	(3) Moderators effects
Full Professor	-0.360 (0.234)	-0.308 (0.225)	-0.432 (0.248)
Associate Professor	-0.316 (0.234)	-0.250 (0.225)	-0.409 (0.248)
Assistant Professor	-0.313 (0.234)	-0.266 (0.225)	-0.379 (0.248)
Postdoctoral researcher	-0.325 (0.234)	-0.273 (0.225)	-0.398 (0.248)
Graduate student	-0.240 (0.242)	-0.218 (0.231)	-0.271 (0.262)
Research Assistant	-0.291 (0.234)	-0.266 (0.225)	-0.327 (0.250)
Constant	0.408* (0.234)	0.368* (0.225)	0.466 (0.248)
Observations	10680	6188	4420
R ²	0.026	0.021	0.034

Note: * $p < 0.05$; ** $p < 0.005$. Standard errors clustered at individual level.

6.3.2.4 Deviation from the pre-analysis plan for the forecasting survey

Below we list three deviations from the pre-registered plan with regard to our forecasting analyses and descriptions of these results.

Testing forecasts about moderators separately by country. As a robustness check for the exploratory hypothesis ‘Do participants predict experimental results across different populations with different degrees of accuracy?’ (estimates presented in Table S6-4), we pre-specified that we would analyze the accuracy of predictions regarding simple effects and regarding moderator effects separately. However, we were mistaken in planning this as the test is in fact impossible. Based on the design of the forecasting survey, the predictions regarding moderators do not vary across countries, since the participants were asked about the effect sizes of the moderators ‘*aggregating across all the replication sites.*’ Since we have no variation in the effect of moderators within different populations, we simply could not run this analysis.

Comparing significance levels of forecasted and replicated effect sizes. We did not pre-register that we would compare whether the forecasted and realized effect sizes for each original effect targeted for replication would respectively differ from zero. However, it can be readily inferred from Table S6-3 where we report the effect sizes and their associated standard errors. It is clear from Table S6-3 that forecasters predicted that all five key effects would be observed (the mean of the forecasts is statistically significantly higher than zero, $p < 0.005$, for all the five key effects). This differs from the realized effect sizes where there was statistically significant support for three of the five key effects: the needles work main effect (works vs. retires comparison), the intuitive work morality effect, and the tacit inferences effect. In contrast, the null hypothesis of no observed effect could not be rejected for the target age and needles work effect and the salvation

prime and work behavior effect (see Table S6-3 for the realized effect sizes and their standard errors).

Splitting forecasted and realized effect sizes by sample. In Table S6-7 below, we report the forecasted and realized effect sizes separately for each sample. This is done for descriptive purposes, without any statistical tests for differences. Note that estimates for “All USA” and “India” are based on Amazon Mechanical Turk (MTurk) samples, whereas the subregions of the USA (New England U.S. states vs. other U.S. states), Australia, and the UK are PureProfile (PP) samples. Data for the salvation prime replication was not collected on MTurk, therefore those entries are blank.

Table 6-7: Forecasted and realized effect sizes separately for each major sample of participants.

		Needless work main effect (works vs. retires)	Target age and needless work effect	Intuitive work morality effect	Tacit inferences effect	Salvation primes and work behavior
New England U.S. States (PP)	Mean Forecast	0.389	0.295	0.292	0.340	0.099
	Actual Effect Size	0.622	0.057	0.313	0.530	0.011
Non New England U.S. States (PP)	Mean Forecast	0.333	0.248	0.244	0.312	0.098
	Actual Effect Size	0.573	0.109	0.247	0.612	-0.031
All USA (MTurk)	Mean Forecast	0.346	0.275	0.259	0.300	-
	Actual Effect Size	0.658	-0.048	0.542	0.641	-
Australia (PP)	Mean Forecast	0.293	0.210	0.246	0.261	0.079
	Actual Effect Size	0.771	0.027	0.329	0.465	-0.107

India (MTurk)	Mean Forecast	0.266	0.200	0.212	0.349	-
	Actual Effect Size	0.666	0.022	-0.067	0.198	-
UK (PP)	Mean Forecast	0.313	0.248	0.258	0.305	0.112
	Actual Effect Size	0.630	0.045	0.175	0.590	0.113

6.4 References

1. T. Andrew P. Ideological inheritance: Implicit Puritanism in American moral cognition. Yale University; 2007
2. Uhlmann EL, Poehlman TA, Tannenbaum D, Bargh JA. Implicit Puritanism in American moral cognition. *J Exp Soc Psychol*. 2011 Mar 1;47(2):312–20.
3. Jost P of PJT. *Social and Psychological Bases of Ideology and System Justification*. Oxford University Press, USA; 2009. 548 p.
4. Camerer CF, Dreber A, Forsell E, Ho T-H, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. *Science*. 2016 Mar 25;351(6280):1433–6.
5. DellaVigna S, Pope D, Vivaldi E. Predict science to improve science. *Science*. 2019 Oct 25;366(6464):428–9.
6. Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, et al. Using prediction markets to estimate the reproducibility of scientific research. *Proc Natl Acad Sci*. 2015 Dec 15;112(50):15343–7.
7. Forsell E, Viganola D, Pfeiffer T, Almenberg J, Wilson B, Chen Y, et al. Predicting replication outcomes in the Many Labs 2 study. *J Econ Psychol*. 2019 Dec 1;75:102117.
8. Landy J, Jia M, Ding I, Viganola D, Tierney W, Dreber A, et al. Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychol Bull*. 2019 Oct 29 [cited 2020 Jan 20]; Available from: <http://repository.essex.ac.uk/25784/>
9. Tierney W, Hardy JH, Ebersole CR, Leavitt K, Viganola D, Clemente EG, et al. Creative destruction in science. *Organ Behav Hum Decis Process*. 2020 Nov 1;161:291–309.
10. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. *Nat Hum Behav*. 2018 Jan;2(1):6.

7 Chapter 7: Forecasting the future of Covid-19 preprints

This chapter consists of the paper “Forecasting the future of Covid-19 preprints” that is currently ready for submission at a journal. I am the main contributor to this study. We employed a similar methodology as was used in SCORE, using prediction markets and surveys to forecast the publication outcomes and future citation counts of 400 preprints. This paper demonstrates that human forecasts can be informative to scientific outcomes beyond experimental results.

Forecasting the future of Covid-19 preprints

7.1 Abstract

The scientific community reacted quickly to the *Covid-19* pandemic in 2020, generating an unprecedented increase in publications. Many of these publications were released on preprint servers such as *medRxiv* and *bioRxiv*. It is however unknown how reliable these pre-prints are, and if they will eventually be published in scientific journals. In this study we use crowdsourced human forecasts to predict publication outcomes and future citation counts for a sample of 400 preprints with high altmetric score. Most of the preprints in this sample were published within one year of upload on a preprint server (69%), and 45% of the published preprints appeared in a high impact journal with Journal Impact Factor of at least 10. On average, the preprints received 161 citations within the first year. We found that forecasters can predict if preprints will be published after one year ($r = 0.22$). Moreover, for published preprints they predict whether the publishing journal has high impact ($r = 0.38$). Forecasts are also informative with respect to *Google Scholar* citations within one year of upload on a preprint server ($r = 0.75$). Subjective assessments of preprints' 'agreement' with other related findings and their helpfulness for mitigating the impacts of the *Covid-19* pandemic are correlated with forecasts and observed outcomes. These forecasts can help to provide a preliminary assessment of preprints at a faster pace than the traditional peer-review process.

7.2 Introduction

The quick rise of the *Covid-19* pandemic was followed by an unprecedented explosion in *Covid-19* related research(1–5). The largest increase in the volume of academic papers from the previous year was in 2020(6) and at the time of writing *PubMed* contained nearly 170,000 *Covid-19* related publications in its database. The dynamics of the pandemic necessitated research findings to be disseminated quickly to other researchers as well as policy and decision makers. Preprint servers helped to accelerate this process by making data and research findings accessible without delays from traditional publications including peer-review(7–9). The fast turnaround was credited to have helped mitigating the potential impacts of the pandemic, including saving lives(1,10). It has been estimated that once the pandemic became widespread in early 2020, 40–50% of all *Covid-19* related publications were submitted to preprint servers before entering the traditional academic publishing route (2,3). *Covid-19* dominated preprint servers, with over half of all preprints on *medRxiv* being *Covid-19* related in every month between March 2020 and August 2021(6). Preprints typically differ from published manuscripts in that they are shorter and contain fewer references(3). In addition, the type of research appearing in preprints differs from traditional publication, with randomised controlled trials (RCTs), systematic reviews, and observational findings appearing more often in preprints, and traditional publications including more case reports and letters(2).

The benefit of a faster dissemination of results comes potentially at the cost of a lower reliability of the released findings(1,4,7,11–13). While benefits and risks of preprints are well documented, a consensus on whether the benefits outweigh the risks has not been reached(14–19). Many preprints are not at ‘publication quality’, and have flaws in data or methods(1). This raises concerns when findings from preprints are shared in traditional and social media, even with the

proviso by preprint servers that the manuscript has not undergone peer-review(20,21). Due to their speed, preprints rather than peer-reviewed publications can be the focus of discourse(4). While people of academic and non-academic backgrounds interact with preprints on social media such as *Twitter*, the difference between peer-reviewed and non-peer-reviewed research may not be understood and flawed studies may be disseminated through media(12,20–22). This problem can be exacerbated by media searching for ‘scoops’ and therefore focusing on exciting and new, but potentially unreliable findings in preprints(8).

It has been estimated that only 5– 14% of *Covid-19* related preprints will be published(3,23,24) and preprints which are eventually published often undergo significant changes, in part due to the peer review process(13). While remaining unpublished for a prolonged time can be an indication of the quality of a preprint, some preprints have been so erroneous in their claims that they have been retracted. The latter includes one infamous preprint which reported that the *Covid-19* virus contained HIV ‘insertions’(1,4). It should, however, be noted that even peer-reviewed published findings cannot be assumed to always be correct or without error(25,26).

In the *Covid-19* preprint forecasting study, we investigate if forecasting can help to fill the gap from preprints lacking peer review. It is currently unknown if crowd-based forecasting allows to identify which preprints are useful and robust and which preprints have flaws which will prevent publication in an academic journal. We asked forecasters to predict publication outcomes and future citations of 400 preprints uploaded on preprint servers between 01 Jan 2020 and 31 August 2020.

The selection of preprints was based on *Altmetric* scores. *Altmetric* scores are an alternate to traditional impact measures such as citations, and are calculated based on metrics such as (but not limited to) mentions on social media, mainstream media coverage, discussion on research blogs and

citations on Wikipedia(27). We split our sample into 10 bins by time and selected the top 40 preprints in each bin as ranked by *Altmetric*. This sampling method was designed so we could test the most widely disseminated preprints across an extended period, which can be expected to include the most relevant findings. We conducted incentivized surveys in November 2020, asking forecasters to predict the probabilities of three possible futures of each preprint: (i) remaining unpublished within one year of dissemination, (ii) being published in a medium or low impact journal, or (iii) being published in a high impact journal. We defined medium or low impact as a journal impact factor (JIF) below 10 and high impact as a JIF of at least 10. In addition, we asked forecasters to predict how a preprint will rank in terms of citations received after one year, relative to the other preprints, with a rank of 0 assigned to the least cited preprint and a rank of 100 to the most cited preprint. We also elicited more subjective assessments such as ‘usefulness’ and replicability. Previous research has shown that forecasters can predict characteristics of papers including their replicability(28–30). While we will not ‘resolve’ the subjective assessments of preprints, these forecasts serve as measure of expectation of their usefulness and replicability.

Our study seeks to understand the extent to which researchers can predict the future impact of preprints through forecasting publication and citation outcomes. Our approach can help to inform future policy around the use of preprints. In addition, our forecasts can act as measure of the quality and usefulness of a manuscript; the nature of preprints provide opportunity for a crowd of informal assessments as opposed more formal assessment through a few peer reviewers and an editor(8).

7.3 Methods

Our sample for contains 400 *Covid-19* preprints from the approximately 6,000+ preprints available at the time of data collection in the *medRxiv* and *bioRxiv Covid-19 SARS-CoV-2*

collection (connect.biorxiv.org/relate/content/181). Of the 400 preprints, 92 were found on *bioRxiv* and 308 were found on *medRxiv*. To focus on preprints that have been recognized by media and social media to be of high relevance, we divided the roughly 6,000 preprints into ten bins by time and selected the top 40 within each bin as ranked by the *Altmetric* score. This stratified sampling strategy results in an even distribution of preprints across time, and avoids the sample being dominated by earlier preprints which had more time to gain a higher *Altmetric* score. Any preprints that had been indicated by *medRxiv* or *bioRxiv* to already be published was excluded from the sampling process.

Participants were recruited predominately from forecasters in the Systematizing Confidence in Open Research and Evidence (SCORE)(30) project, which focused on forecasting the replicability of scientific claims in the social and behavioural sciences. We had 49 participants in total. The median preprints forecasted by participants is 30 ($M = 93$, $SD = 133$). The participants are typically but not necessarily from academia or have academic backgrounds. Since the SCORE project focused on claims from the social and behavioral sciences, the participants typically had little expertise in biomedical research. Yet, a considerable fraction of researchers had contributed to replication studies and related ‘metascience’ projects. Participants were also recruited via social media, primarily *Twitter* and *Reddit*. Participants were assigned an initial random ‘batch’ of 10 preprints; once the initial batch was completed, participants had the option to complete additional batches. The surveys were open from October 28, 2020, through November 10, 2020. Participants had no access to other participants forecasts. The abstract of the preprints was provided within the survey along with links to full version of the online preprint. The surveys included four forecasting questions (with answers or prompts).

(Q1) Will this preprint be published in a peer-reviewed scientific journal within a year of first preprint posting? Provide a % probability between 0 and 100 for each option. The values must

sum to 100: (option 1) No, not published, (option 2) Yes, in a journal with impact factor below 10, (option 2) Yes, in a journal with impact factor of at least 10.

(Q2) Rank this preprint's one-year Google citations count relative to other preprints in this study. Select (using slider) a relative rank between 0 for least cited and 100 for most cited).

(Q3) What is the % probability that the findings presented in the preprint agree with the majority of results from similar future studies? Select (using slider) a value between 0 (impossible) and 100 (certain).

(Q4) Are the results presented in the preprint helpful to mitigate the impact of the COVID pandemic? Select (using slider) a value between 0 (no) and 100 (yes).

Because experiences from previous projects suggest that participants often provide 'conflated' responses when asked about different aspects of a publication, we asked Q3 and Q4 in random order.

Incentives for the surveys were provided through surrogate scoring (31). This method does not require access to a 'ground truth' outcome to incentivise truthful reporting and is thus well-suited to elicit a broader range of judgements. It is also well-suited to generate accuracy estimates for paying prizes without delay if resolution is not available immediately, such as in this case. The core idea of the surrogate scoring approach is to first identify a surrogate outcome using the collected judgement (the mean). Then we develop a statistical estimation procedure to uncover the bias in this surrogate and noisy outcome. This knowledge of the bias helps us define unbiased estimates of the true scores as if we had access to the ground truth outcome, using only the surrogate outcome(31). Prizes for the surveys were given for the best forecasters in each batch.

With 40 batches of 10 questions, we paid USD 90 per batch, paid as \$30, \$25, \$20, and \$15 to the 1st, 2nd, 3rd, and 4th place for that batch, as determined by the surrogate scoring method.

Participants could participate and win prizes in more than one batch.

The questions Q1 and Q2 have been resolved one year after the preprint was uploaded to a preprint server. Publication outcomes were resolved manually using *Google Scholar* searches for any published version of a particular preprint. Resolutions allowed for some differences in preprints and publications such as updated sample sizes or edited text. In cases where published articles were made available online before the official journal edition publication date, the earliest date was used as the publication date. The journal impact factor (JIF) in 2020 was used for resolving published preprints. The citation counts were also resolved manually using the total number of citations recorded by *Google Scholar*, summing-up citation counts of multiple instances of the paper (including preprints and published version) being indexed by *Google Scholar*. Where preprints were not resolved after one year, citation counts were backdated using *Google Scholar*'s citations by date function (this was applied to 86 preprints). Q3 and Q4 are not resolved but serve as a gauge of the expectations of the forecasters of the robustness and helpfulness (regarding the mitigation of the pandemic) of the preprints. We also conducted a prediction market to forecast above outcomes but decided not to use the corresponding data due to technical glitches. Problems included that starting prices were set incorrectly, and prices on the dashboard did not update correctly. The combination of these technical issues, coupled with a low participation of only around 30 traders led us to not analyse the prediction markets, although the unused data will be made available. This decision was made before any data were analysed.

The experimental design and statistical analyses were pre-registered on the *Open Science Framework* (OSF) after the surveys were completed, towards the end of the data collection for the resolutions. We did not use the surveys or incomplete outcome data for any preliminary analyses.

Any deviations from the pre-registered analyses or any added not pre-registered tests will be mentioned in the text below. For statistical tests, we interpret the threshold of $p < 0.005$ as identifying statistical significance, and the threshold of $p < 0.05$ as identifying suggestive evidence(32).

7.4 Results

Of the 400 preprints in our sample, 276 (69%) were published in an academic journal within one year. The median time between the release date on a preprint server and becoming published is 143 days ($M = 152$, $SD = 81$, $Min = 14$, $Max = 362$). Note that more preprints in our sample may have been published or will be published outside our one-year cutoff. 146 of the 276 (45%) published papers were published in a high impact journal (JIF of at least 10). Papers were published in 135 different journals with the most common journals being *Nature* (22 papers), *Science* (16 papers) and *Nature Communications* (14 papers).

We also collected citation counts for each of the 400 preprints, combining *Google Scholar* citation counts for all versions including preprints and published versions (where relevant). The median number of citations after one year (from the release date on a preprint server) was 54 ($M = 161$, $SD = 307$, $Min = 14$, $Max = 3,286$). 54 papers had less than 10 citations, 200 papers had between 10 and 100 citations, 138 papers between 100 and 1000, and 8 papers had more than 1000 citations. We transform citation counts into relative ranks between 0 and 100, where a paper with rank 0 has the lowest citation count and a paper with rank 100 has the highest citation count. These ranks are used in our statistical analyses (instead of the actual citation counts) with the advantage of avoiding inferential problems that may arise due to the highly non-normal (right-skewed) distribution of citation count data.

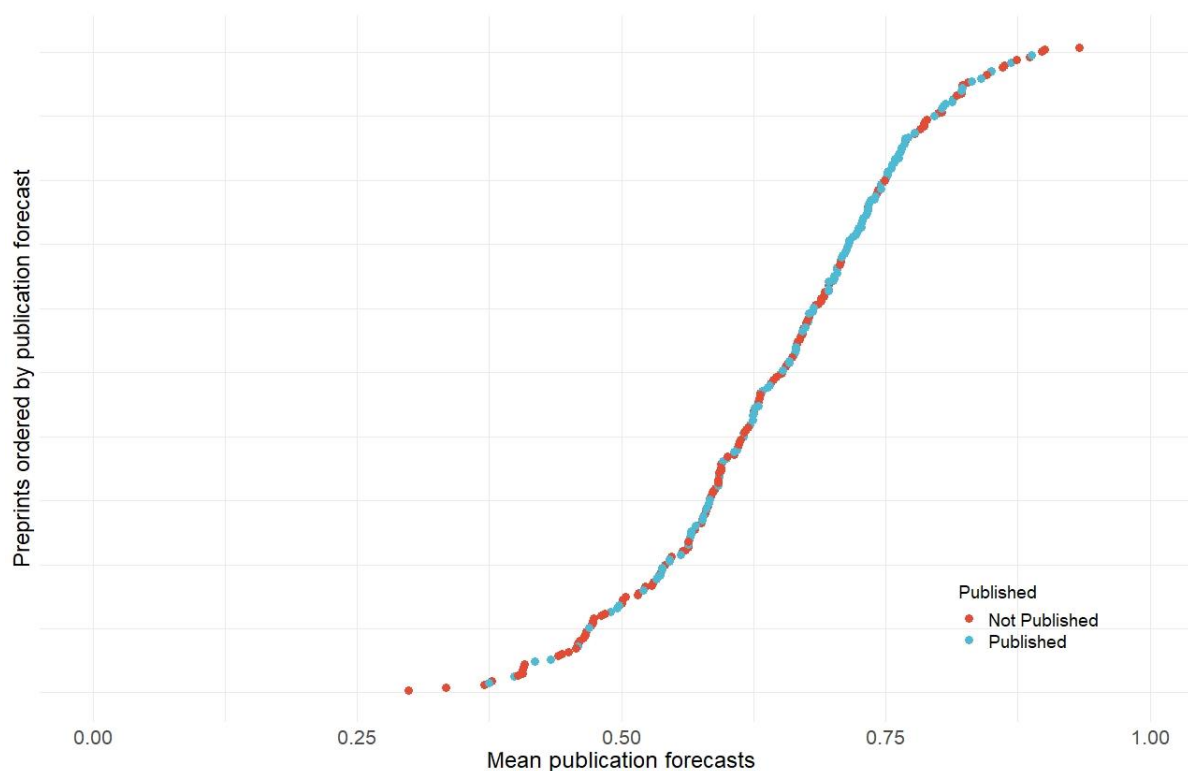
Publication status (i.e., whether or not the paper has been published) is statistically significantly correlated with the paper's citation rank ($r = 0.40$, $t(398) = 8.64$, $p < 0.0001$), with non-published preprints having an average citation rank of 33 compared with published preprints' average rank of 58, a statistically significant difference (un-preregistered Welch's t -test, $t(273.12) = 9.15$, $p < .0001$). For the preprints which have been published, citation ranks are statistically significantly correlated with the impact factor category ($r = 0.50$, $t(268) = 9.36$, $p < 0.0001$) where impact factor category refers to a binary variable indicating publication in a high impact or medium to low impact journal. Preprints published in journals with impact factors of at least 10 have an average citation rank of 73 compared with an average citation rank of 45 for preprints published in journals with a JIF below 10 (un-preregistered Welch's t -test, $t(265.65) = 9.42$, $p < 0.0001$). Preprints published in low to medium impact journals (JIF < 10) are cited statistically significantly more than non-published preprints (un-preregistered Welch's t -test, $t(264.01) = -4.16$, $p < 0.0001$).

Due to the delay between sampling preprints and conducting the surveys and as publication flags on preprint servers are not comprehensive, 148 preprints had already been published by the end of the survey period (10 November 2020). Therefore, these papers have been excluded when assessing the performance of the forecasters in predicting publication outcome, leaving 128 published preprints (published after 10 November 2020) and 124 non-published preprints.

The mean forecasts for publication status (combining forecasts for publication in any journal in this instance) are statistically significantly positively correlated with the binary publication outcome (where 0 = not published, 1 = published) ($r = 0.22$, $t(250) = 3.49$, $p = 0.0006$) (see Figure 1). We also regressed individual responses against publication outcomes in a linear probability model with standard errors clustered at the forecaster level. This model showed suggestive evidence for a positive association between individual forecasts for publication in any

journal and the binary publication outcome ($\beta = 0.15$, $t(47) = 2.36$, $p = 0.018$). Forecasters statistically significantly overestimate the overall publication rate when comparing binary publication outcomes and average forecasts of probability of publication: forecasters expect 65% of preprints to be published compared to the reality of 51% (paired t -test, $t(251) = 4.46$, $p < 0.0001$). When assessing binary accuracy, where we interpret an average survey forecast of less than 0.5 as prediction of a preprint not being published, and an average survey forecast of 0.5 or above as a prediction of a preprint being published, forecasters have an accuracy of 58%. There is suggestive evidence that this accuracy rate is better than random chance (un-preregistered binomial test, $p = 0.01958$)

Figure 7-1 – Forecasts of publication outcomes. This figure plots the mean forecasted probability of a preprint becoming published with one year, ordered by the mean forecast. The color of the markers indicates the publication status (irrespective of the impact factor of the journal), with blue markers indicating published preprints and red markers indicating preprints not published within one year. Publication outcomes are correlated with mean forecasts ($r = 0.22$, $t(250) = 3.49$, $p < 0.001$). Preprints for which the publication status was known during the forecasting period are excluded from this figure.



We also tested if forecasters could predict the quality of the journal the preprint will be published in. We find a statistically significant correlation between the mean survey predictions for being published in a high impact journal (given that it is published) and the binary outcome, where 0 = published in low-medium impact journal and 1 = published in a high impact journal ($r = 0.38$, $t(122) = 4.58$, $p < 0.0001$) (see Figure 2). In addition, a linear probability model was used on the individual level responses with clustered standard errors at the forecaster level. This model showed positive association between forecasts and outcomes ($\beta = 0.27$, $t(47) = 4.32$, $p < 0.0001$). There was suggestive evidence that forecasters tended to underestimate the share of papers which were published in high impact journals (forecast average of 27%) compared with 36% in reality (paired t -test, $t(123) = -2.22$, $p = 0.029$).

Forecasters can predict which preprints will have few or many citation counts after one year. Citation ranks (from 0 to 100) are statistically significantly correlated with the mean forecasts for citation ranks ($r = 0.75$, $t(398) = 22.36$, $p < 0.0001$). The strength of this relation is illustrated in Figure 3. As the preprints in our sample were released at different times, we test if papers which were released early, and therefore had more time to get indications of citations, have a systematically lower absolute error. We correlate days between upload to preprint server and the end of the survey period (10 of November 2020) and absolute citation rank error in an unpreregistered test. We find no evidence that being released earlier correlates with more accurate forecasts ($r = -0.10$, $t(398) = -1.95$, $p = 0.052$).

Figure 7-2 – Forecasts of journals of publication. The figure shows the mean forecasted probability of a preprint being published with one year, given that a preprint is published, ordered by mean probability. The color of the markers indicates if the preprints have been published in a high impact journal, with blue markers indicating publication in a high impact journal and red markers indicating preprints published in a medium or low impact journal. Journal publication outcomes are correlated with the mean forecasts ($r = 0.38$, $t(122) = 4.58$, $p < 0.001$).

Unpublished preprints and preprints for which the publication status was known during the forecasting period are excluded from this figure.

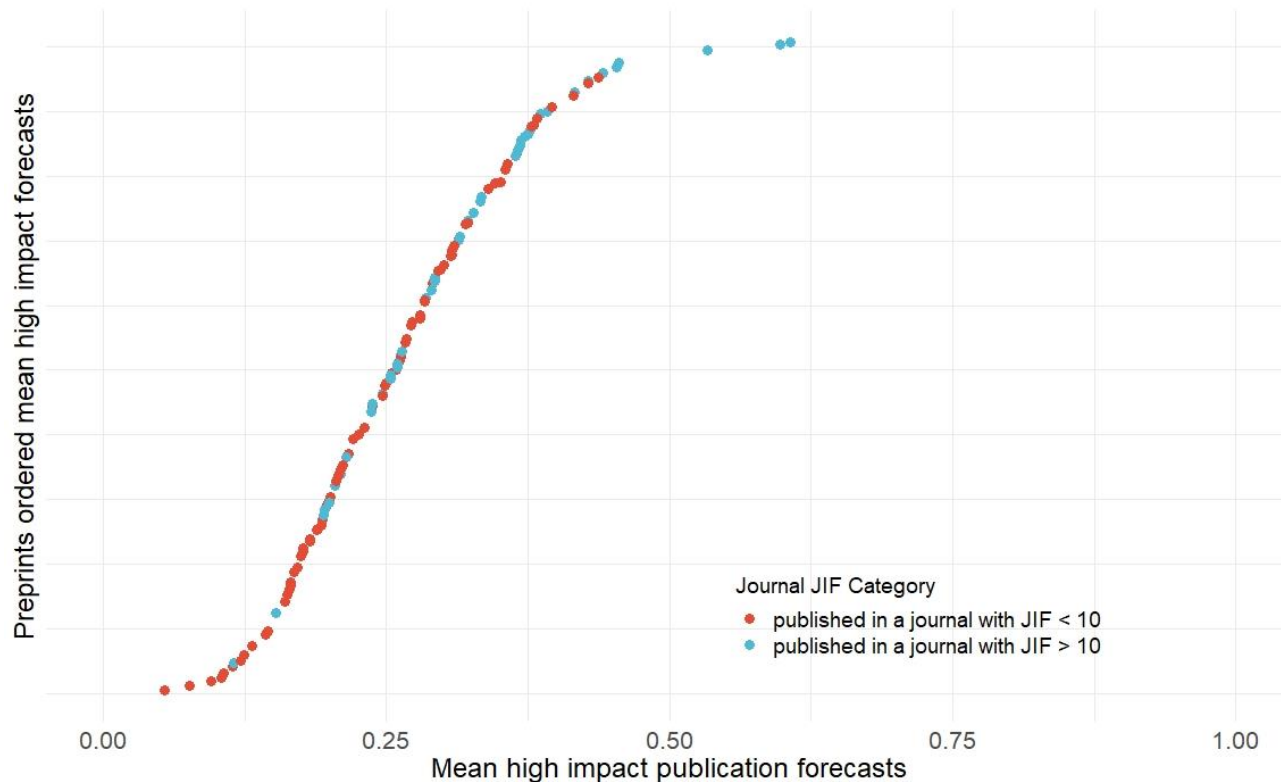
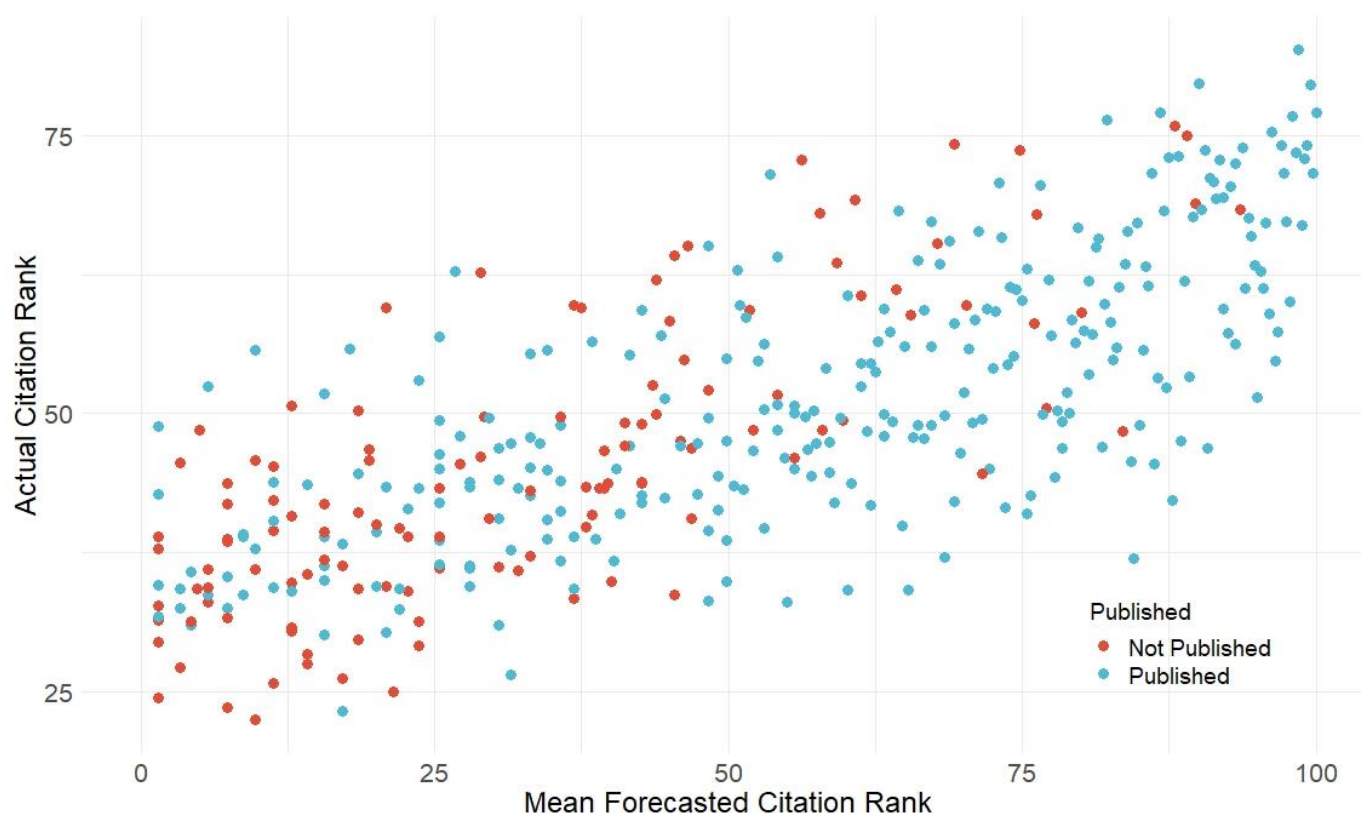


Figure 7-3 – Actual citation ranks vs. forecasted citation ranks. This plot demonstrates the relationship between mean forecasted citation ranks and actual citation ranks. The color of the markers indicates the publication status (irrespective of the impact factor of the journal), with blue markers indicating published preprints and red markers indicating preprints not published within one year. While the mean forecasts are highly correlated with realised citation ranks ($r = 0.75$, $t(398) = 22.36$, $p < 0.001$), the aggregated forecasts are not extreme enough with few preprints forecasted to be ranked below 25 or above 75.



The survey responses were incentivised using SSR which provides estimates of a forecasters AUC or Brier score. We tested the accuracy of SSR in identifying the most accurate forecasters. Using SSR we estimated the rank of users by AUC and Brier score in each batch (a batch is made up of 10 preprints) and compare the estimated ranks with actuals. We find that estimated ranks of batchwise AUC and Brier score are correlated with actual ranks (AUC: $r = 0.16$, $t(471) = 3.55$, $p = .0004$, Brier Score: ($r = 0.17$, $t(471) = 3.66$, $p = .0003$)). We also correlated overall users SSR estimated AUC and Brier Scores with actual AUC and Brier Scores. We find that overall SSR estimates are not correlated with actual AUC ($r = -0.10$, $t(30) = -0.54$, $p = 0.5954$) or Brier Score ($r = 0.29$, $t(30) = 1.67$, $p = 0.1057$). SSR can also be used to aggregate forecasts by taking an average of only the forecasters with highest estimated AUC, with the aim to remove uninformative forecasts. We find evidence that squared error of SSR aggregated citation rank and is

higher than the squared error of the mean aggregated citation rank (paired t -test, $t(399) = -10.63$, $p < .0001$).

In addition to the resolvable questions of publication and citation outcomes, we also ask forecasters to provide subjective judgements on the replicability and usefulness (with respect to the mitigation of the pandemic) of the preprints. While not being able to test for accuracy of these forecasts, they serve as a gauge of the expectations of how the findings in our sample sit in the wider body of *Covid-19* literature. We find that average responses to Q3 (What is the % probability that the findings presented in the preprint agree with the majority of results from similar future studies?) correlate positively with forecasts for publication ($r = 0.65$, $t(250) = 13.39$, $p < .0001$), forecasted citation rank ($r = 0.52$, $t(398) = 12.02$, $p < .0001$) and forecasted usefulness ($r = 0.62$, $t(398) = 15.74$, $p < .0001$). In terms of outcomes, forecasts for agreement (Q3) are also positively correlated with being published, being published in a high impact journal, and the citation rank ($r = 0.35$, $t(398) = 7.39$, $p < .001$; $r = 0.25$, $t(268) = 4.23$, $p < .0001$; and $r = 0.43$, 95% CI [0.34, 0.50], $t(398) = 9.38$, $p < .0001$; respectively). The assessed usefulness of preprints (as measured by Q4) is also correlated with forecasts of being published ($r = 0.70$, $t(250) = 15.68$, $p < .001$). Forecasters expect that more useful papers are more likely to be published in high impact journals: responses to Q4 are more correlated to forecasts of being published in a high impact journal than being published in a low-medium impact journal ($r = 0.65$, $t(250) = 13.65$, $p < .001$ vs. $r = 0.35$, $t(250) = 5.92$, $p < .001$ and statistically significant difference $p < .0001$). Forecasts of usefulness is also positively related to publication in a high impact journal ($r = 0.33$, $t(268) = 5.81$, $p < .001$) and with actual citation ranks ($r = 0.50$, $t(398) = 11.47$, $p < .001$).

As participants may provide conflated answers for Q3 and Q4, we randomised the order these questions with of half the participants being asked Q3 first; the other half Q4 first. To test for order effects, we compare responses to Q3 and Q4 between those who answered Q3 or Q4 first

using unpaired t -tests. We find that responses are significantly lower when Q3 is asked first as opposed to asked second (unpaired t -test, difference = -0.07 , $t(545.26) = -7.10$, $p < 0.001$). No such order effects are found for Q4 (unpaired t -test, difference = -0.01 , $t(558.70) = -0.90$, $p = 0.370$). All correlations including Q3 split by order of asking can be found in Table 1.

Table 7-1 – Pairwise Pearson correlation coefficients of survey questions and outcomes. The order of questions 3 and 4 were randomised so that some participants always saw question 3 first and some always saw question 4 first. We found evidence for order effects of for question 3 and so include all answers for question 3 (labelled “Q3 - Agreement with other papers” in the table) and split by whether it was asked first or second.

	Q1 - Not published	Q1 - Published (IF <10)	Q1 - Published (IF > 10)	Combined published forecast (any IF)	Q2 - Cite Rank	Q3 - Agreement with other papers	Q3 - Agreement with other papers (asked first)	Q3 - Agreement with other papers (asked second)	Q4 - Helpful	Publication Outcome	Published (IF>10)
Q1 - Published (IF <10)	-0.7 (p <.0001)										
Q1 - Published (IF > 10)	-0.74 (p <.0001)	0.03 (p =0.61997)									
Combined published forecast (any IF)	-1 (p <.0001)	0.7 (p <.0001)	0.74 (p <.0001)								
Q2 - Cite Rank	-0.78 (p <.0001)	0.43 (p <.0001)	0.69 (p <.0001)	0.81 (p <.0001)							
Q3 - Agreement with other papers	-0.65 (p <.0001)	0.5 (p <.0001)	0.43 (p <.0001)	0.64 (p <.0001)	0.52 (p <.0001)						
Q3 - Agreement with other papers (asked first)	-0.53 (p <.0001)	0.38 (p <.0001)	0.38 (p <.0001)	0.56 (p <.0001)	0.5 (p <.0001)	0.9 (p <.0001)					
Q3 - Agreement with other papers (asked second)	-0.56 (p <.0001)	0.45 (p <.0001)	0.35 (p <.0001)	0.53 (p <.0001)	0.35 (p <.0001)	0.76 (p <.0001)	0.46 (p <.0001)				
Q4 - Helpful	-0.7 (p <.0001)	0.35 (p <.0001)	0.65 (p <.0001)	0.7 (p <.0001)	0.69 (p <.0001)	0.62 (p <.0001)	0.56 (p <.0001)	0.48 (p <.0001)			
Publication Outcome	-0.21 (p =0.00068)	0.2 (p =0.00116)	0.1 (p =0.09813)	0.34 (p <.0001)	0.24 (p <.0001)	0.35 (p <.0001)	0.31 (p <.0001)	0.32 (p <.0001)	0.23 (p <.0001)		
Published (IF>10)	-0.22 (p =0.00043)	0.01 (p =0.81874)	0.29 (p <.0001)	0.34 (p <.0001)	0.4 (p <.0001)	0.3 (p <.0001)	0.29 (p <.0001)	0.23 (p <.0001)	0.33 (p <.0001)	0.45 (p <.0001)	
Actual Cite Rank	-0.55 (p <.0001)	0.3 (p <.0001)	0.49 (p <.0001)	0.62 (p <.0001)	0.75 (p <.0001)	0.43 (p <.0001)	0.43 (p <.0001)	0.27 (p <.0001)	0.5 (p <.0001)	0.4 (p <.0001)	0.53 (p <.0001)

7.5 Discussion

The forecasting of scientific outcomes can help to improve science(28,29,33–36). In previous projects it has been shown that forecasters can predict replicability(28) and effect sizes(37,38). While earlier studies used forecasting to predict the rating of publications in review

exercises such as the REF(39), this is the first project to forecast outcomes pertinent to specific preprints. The use and function of preprints has changed over the course of the global *Covid-19* pandemic – providing potentially lifesaving findings and data to policy makers(1,10). The undeniable benefits of preprints do come at a cost – the lack of peer-review can result in some preprints lacking credibility. We sought to understand the extent to which forecasters can differentiate the credibility of preprints by forecasting publication and citation outcomes. We elicited through incentivised surveys; forecasted probability of publication in a peer-reviewed journal (with 3 outcomes including impact factor of journal), forecasted relative citation rank, and two non-verifiable characteristics of the preprint – the agreement with other results (i.e., replicability) and the usefulness of mitigating the impact of the pandemic.

Our results showed that forecasters can predict if preprints will be published within one year, despite overestimating the share of preprints in our sample which will be published. Forecasters were also able to predict the impact of journal in which the preprints will be published, despite underestimating the number of the papers which are published in high impact journals. Similarly, our forecasts of relative citation ranks were also highly correlated with actual relative citation ranks. We also found that subjective assessments of replicability and usefulness are correlated with publication outcomes and citation ranks. Our forecasters were more accurate at predicting citation rank than publication outcomes. We excluded a number of preprints from our accuracy analysis due them being already published by the time of the forecasting survey. Preprints which are published quickly maybe more obviously ‘publishable’ and therefore are easier to forecast. Preprints which are delayed in being published and therefore are in our sample may be more difficult to forecast. In addition, there is a delay between the prints in our sample being uploaded to a preprint server and the forecasting survey. This means that participants had some information of citations when forecasting. We found that having more information on citations

(from preprint being uploaded earlier from the surveys) was not correlated with more accurate predictions of final citation ranks. Finally, it may be that citations are a more fine-grained and continuous outcome than publication and therefore easier to forecast. Further study with forecasting happening shortly after preprints are released may be used to test these assumptions.

Our findings provide a promising proof of concept that forecasting can be used to help filling the gap left by the lack of peer-review on preprints. Using forecasting can help provide initial signals of credibility and usefulness for both other researchers and for a more general audience when exposed via social or traditional media.

7.6 References

1. Dinis-Oliveira RJ. COVID-19 research: pandemic versus “paperdemic”, integrity, values and risks of the “speed science.” *Forensic Sci Res.* 2020 Apr 2;5(2):174–87.
2. Gianola S, Jesus TS, Barger S, Castellini G. Characteristics of academic publications, preprints, and registered clinical trials on the COVID-19 pandemic. *PLOS ONE.* 2020 Jun 10;15(10):e0240123.
3. Lachapelle F. COVID-19 Preprints and Their Publishing Rate: An Improved Method. *medRxiv.* 2020 Sep 7;2020.09.04.20188771.
4. Majumder MS, Mandl KD. Early in the epidemic: impact of preprints on global discourse about COVID-19 transmissibility. *Lancet Glob Health.* 2020 May 1;8(5):e627–30.
5. Vasconcelos GL, Cordeiro LP, Duarte-Filho GC, Brum AA. Modeling the Epidemic Growth of Preprints on COVID-19 and SARS-CoV-2. *Front Phys.* 2021;9:125.
6. Brainard J. A COVID-19 publishing revolution? Not yet. *Science.* 2021;373(6560):1182–3.
7. Guterman EL, Braunstein LZ. Preprints During the COVID-19 Pandemic: Public Health Emergencies and Medical Literature. *J Hosp Med* 2020 Oct 1 [cited 2021 Sep 15];15(10). Available from: <https://www.journalofhospitalmedicine.com/jhospmed/article/228330/hospital-medicine/preprints-during-covid-19-pandemic-public-health>
8. Hoy MB. Rise of the Rxivs: How Preprint Servers are Changing the Publishing Process. *Med Ref Serv Q.* 2020 Jan 2;39(1):84–9.
9. Kleinert S, Horton R. Preprints with The Lancet are here to stay. *The Lancet.* 2020 Sep 19;396(10254):805.
10. Besançon L, Peiffer-Smadja N, Segalas C, Jiang H, Masuzzo P, Smout C, et al. Open science saves lives: lessons from the COVID-19 pandemic. *BMC Med Res Methodol.* 2021 Jun 5;21(1):117.
11. Kirkham JJ, Penfold NC, Murphy F, Boutron I, Ioannidis JP, Polka J, et al. Systematic examination of preprint platforms for use in the medical and biomedical sciences setting. *BMJ Open.* 2020 Dec 1;10(12):e041849.

12. Bagdasarian N, Cross GB, Fisher D. Rapid publications risk the integrity of science in the era of COVID-19. *BMC Med.* 2020 Jun 25;18(1):192.
13. Carneiro CFD, Queiroz VGS, Moulin TC, Carvalho CAM, Haas CB, Rayêe D, et al. Comparing quality of reporting between preprints and peer-reviewed articles in the biomedical literature. *Res Integr Peer Rev.* 2020 Dec 1;5(1):16.
14. Flanagin A, Fontanarosa PB, Bauchner H. Preprints Involving Medical Research—Do the Benefits Outweigh the Challenges? *JAMA.* 2020 Nov 10;324(18):1840–3.
15. Schalkwyk MCI van, Hird TR, Maani N, Petticrew M, Gilmore AB. The perils of preprints. *BMJ.* 2020 Aug 17;370:m3111.
16. Gopalakrishna G. Preprint advocates must also fight for research integrity. *Nature.* 2021 Sep 13 [cited 2021 Sep 16]; Available from: <https://www.nature.com/articles/d41586-021-02481-y>
17. Sheldon T. Preprints could promote confusion and distortion. *Nature.* 2018 Jul 24;559(7715):445–445.
18. Sarabipour S. Preprints are good for science and good for the public. *Nature.* 2018 Aug 29;560(7720):553–553.
19. Soderberg CK, Errington TM, Nosek BA. Credibility of preprints: an interdisciplinary survey of researchers. *R Soc Open Sci.* 7(10):201520.
20. Carlson J, Harris K. Quantifying and contextualizing the impact of bioRxiv preprints through automated social media audience segmentation. *PLOS Biol.* 2020 Sep 22;18(9):e3000860.
21. Ravinetto R, Caillet C, Zaman MH, Singh JA, Guerin PJ, Ahmad A, et al. Preprints in times of COVID19: the time is ripe for agreeing on terminology and good practices. *BMC Med Ethics.* 2021 Jul 28;22(1):106.
22. Glasziou PP, Sanders S, Hoffmann T. Waste in covid-19 research. *BMJ.* 2020 May 12;369:m1847.
23. Fraser N, Brierley L, Dey G, Polka JK, Pálffy M, Nanni F, et al. The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLOS Biol.* 2021 Feb 4;19(4):e3000959.
24. Añazco D, Nicolalde B, Espinosa I, Camacho J, Mushtaq M, Gimenez J, et al. Publication rate and citation counts for preprints released during the COVID-19 pandemic: the good, the bad and the ugly. *PeerJ.* 2021 Mar 3;9:e10927.
25. Ioannidis J. Why Most Published Research Findings Are False. *PLOS Med.* 2005 Aug 30;2(8):e124.
26. Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol Sci.* 2011 Nov 1;22(11):1359–66.
27. Adie E, Roe W. Altmetric: enriching scholarly content with article-level discussion and metrics. *Learn Publ.* 2013;26(1):11–7.
28. Gordon M, Viganola D, Dreber A, Johannesson M, Pfeiffer T. Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. *PLOS ONE.* 2021 Apr 14;16(4):e0248780.
29. Gordon M, Viganola D, Bishop M, Chen Y, Dreber A, Goldfedder B, et al. Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *R Soc Open Sci.* 2020 Jul 22;7(7):200566.
30. Alipourfard N, Arendt B, Benjamin DM, Benkler N, Bishop M, Burstein M, et al. Systematizing Confidence in Open Research and Evidence (SCORE) *SocArXiv*; 2021 [cited 2021 Jun 14]. Available from: <https://osf.io/preprints/socarxiv/46mnb/>

31. Liu Y, Wang J, Chen Y. Surrogate Scoring Rules. *ACM Conf Econ Comput* 2020 Jul; Available from: <http://arxiv.org/abs/1802.09158>
32. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. *Nat Hum Behav*. 2018 Jan;2(1):6.
33. Hanson R. Could gambling save science? Encouraging an honest consensus. *Soc Epistemol*. 1995 Jan 1;9(1):3–33.
34. DellaVigna S, Pope D, Vivaldi E. Predict science to improve science. *Science*. 2019 Oct 25;366(6464):428–9.
35. Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, et al. Using prediction markets to estimate the reproducibility of scientific research. *Proc Natl Acad Sci*. 2015 Dec 15;112(50):15343–7.
36. Landy J, Jia M, Ding I, Viganola D, Tierney W, Dreber A, et al. Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychol Bull*. 2019 Oct 29 [cited 2020 Jan 20]; Available from: <http://repository.essex.ac.uk/25784/>
37. Tierney W, Hardy J, Ebersole CR, Viganola D, Clemente EG, Gordon M, et al. A creative destruction approach to replication: Implicit work and sex morality across cultures. *J Exp Soc Psychol*. 2021 Mar 1;93:104060.
38. Tierney W, Hardy JH, Ebersole CR, Leavitt K, Viganola D, Clemente EG, et al. Creative destruction in science. *Organ Behav Hum Decis Process*. 2020 Nov 1;161:291–309.
39. Munafo MR, Pfeiffer T, Altmejd A, Heikensten E, Almenberg J, Bird A, et al. Using prediction markets to forecast research evaluations. *R Soc Open Sci*. 2(10):150287.

8 Chapter 8: Conclusion

This final chapter contains a summary of the findings, contributions to knowledge and implications of this thesis, followed by potential areas of further study, and finally my final thoughts

8.1 Findings, implications, and contributions to knowledge

The world needs science to be credible. However many scientists' trust in science have been degraded (1,2). Questions regarding the credibility of scientific claims have been raised in several fields across science (3–25). Empirical evidence that not all claims in scientific publications are credible has been provided by large scale replication projects in biology (26), psychology (27–32), experimental economics (33,34), philosophy (35), and the social sciences more generally (36).

Four of these replication projects were accompanied by forecasting projects which utilised crowdsourcing to predict which claims will be confirmed under direct replication and which ones will not. Forecasts that are informative in respect to the replicability of claims can be used to allocate replication resources, act as post publication assessments, and serve to inform the public and other researchers about the expected credibility. A meta-analysis of data from these four forecasting projects, as presented in chapter 2, demonstrates that prediction markets and, to a lesser extent, surveys with simple aggregation, can be used to elicit and aggregate forecasts to predict replicability. Surveys could correctly predict the replicability 66% of the time, compared to 73% for prediction markets. Prediction markets typically provide more extreme forecasts which often provide lower errors which may have contributed to the prediction markets outperforming the surveys. In addition, prediction markets allow forecasters to update their beliefs based on new information including the forecasts of others. The forecasters when filling in the surveys in this

project are not rewarded for accuracy and as such, are not incentivised to put effort into forecasts or truthfully report beliefs, which may contribute to lower accuracy. While unincentivized surveys are cheaper than prediction markets (in terms of incentives only, the other fixed costs for running the project remain similar), the greater predictive power of prediction markets mean they provide much greater benefit for slightly lower cost. The lack of like-for-like comparison (i.e incentivised surveys and incentivised prediction markets) does present a limitation of this study which is addressed in SCORE project – where both surveys and prediction markets are incentivised. This will allow for a more effective comparison between the predictive power of surveys and prediction markets. In addition, the SCORE project also includes elicitation protocols which incentivise participation as opposed accuracy, known as the Delphi protocol. This protocol is implemented by another team at the University of Melbourne.

I demonstrate that the market prices converged such that after an average of 69 trades (or a week after the market opens) price changes added only noise and did not further reduce the forecasting error. Therefore, any replication forecasting project can use prediction markets to provide accurate estimates of replicability. However, there must be sufficient traders and liquidity in the prediction markets to ensure that markets will converge.

The relationship between replication outcomes and p-values of original findings was shown to be robust in the combined data, with 74% of original findings with p-values below 0.005 being replicated, compared to 28% for findings with p-values above 0.005. A relationship between p-values and replicability has been discussed previously (32,48), however here I presented empirical evidence of the relationship between p-values and replicability across fields of study (psychology, economics and general social science) and replication projects. This information confirms the recommendations by Benjamin and colleagues (37), that p-values between 0.05 and 0.005 should be interpreted as ‘suggestive evidence’. In addition, while an uninformative prior is often set as the

starting price in markets (i.e. 0.5) the p-value of the original claim can be used as a prior to set the starting price. Note that these contributions to knowledge have already been applied in the SCORE project.

Finally, my analysis supports previous assertions (38–40) that the mean of individual predictions can be a poor aggregator of probabilistic predictions, and I accordingly provide alternative aggregation methods which outperformed the mean as an aggregator.

These findings however are limited to sample size of around 100⁷, and of extremely similar experimental designs. It is not clear how these results will generalize to other projects with methodological differences. Each of the projects included in the meta-analysis, were small, highly focused forecasting projects with many traders. These findings may not extend to a project like SCORE which is much wider in scale (3000 claims) and scope (multiple academic fields). While prediction markets were shown here to be the most accurate forecasting tool, they were only compared to one other elicitation method – unincentivized surveys. There are however other forms of elicitation, such as the delphi group elicitation methods which could also be included for comparison.

The DARPA SCORE project is the successor of previous smaller forecasting studies with unique scale and scope (41). While previous studies sought to forecast at a scale of about 25 findings, SCORE aims to elicit forecasts for the replicability of the 3000 claims in the social and behavioural sciences, using both human and machine forecasts. The full description of the SCORE project can be found in appendix 1. As part of this large international collaboration, our team used

⁷ Although this sample sizes represents a 4 times larger sample than other similar projects

incentivised prediction markets and surveys to elicit forecasts on the replicability of claims over 12 monthly rounds.

The first monthly round focused on eliciting ‘meta’ forecasts consisting of overall replication rates in SCORE, average rates of replicability across fields of science and average rates of replicability years. The results of this round were presented in chapter 4. We found that forecasters expect replication rates to differ by field, with economics being the field with highest expected replication rate (58%). Psychology and Education have the lowest expected replication rates (42%). While large scale replication projects have been conducted for psychology (32) and economics (33), there are however many fields in which no large-scale replication project can be used to estimate the field-wide replication rate. I therefore provide a noisy estimate (given the accuracy of forecasters in the past (42)) of field wide replication rates in a number of fields where no large-scale replication projects exist. These fields include political science (expected replication rate of 0.49), education (expected replication rate of 0.42) sociology and criminology (expected replication rate of 0.45) and marketing and management (expected replication rate of 0.43). These rates can serve as an informative prior until replication rates can be confirmed via empirical evidence. I also demonstrated that forecasters expect replication rates to increase year on year (2009-2018), showing confidence in the rise of open science practices such as pre-registration (5). The findings from the field wide and yearly replication rates forecasts have the obvious limitation that there is no ground truth to validate the forecasts. Once the SCORE project is complete, there will be the opportunity to assess the accuracy of our forecasts. Validation will also allow for testing of new hypotheses and new analyses including how participant’s characteristics effect forecasts and accuracies. In particular, we will be able to test if participants are more accurate when assessing their own fields, as opposed to out of field forecasts. In addition, this research is limited by the number of participants. With a sample size of 226, this sample may not be representative of the

academic community in terms of community expectations about replication rates – especially in fields outside of economics and psychology which were much less represented in the sample. From a forecasting point of view, this sample size may not be sufficient to provide accurate forecasts or allow the prediction markets to become efficient.

Forecasts for specific claims were elicited over 11 monthly rounds following the initial ‘meta round’ for a total of 12 rounds. Rounds 2-11 assessed 300 social and behavioural claims using surveys and prediction markets for a total of 3000 claims. The final round focused on social and behavioural claims related to COVID-19 and was an addition to the SCORE programme. With replications being conducted by other teams in the SCORE collaboration, and delayed because of the COVID-19 pandemic, to date, results are not available to validate the forecasts from rounds 2 to 12, however the pre-registered analysis plans are shown in chapter 5. This pre-registration includes descriptions of experimental design and methodology and detailed explanations of the ‘confidence scores’ generated for this project, which are our predictions of replicability for the 3,000 SCORE claims. Each confidence score is calculated based on a unique elicitation or aggregation method. Many of these confidence scores are informed by previous conclusions about the aggregation of probabilistic forecasts (42,43), including those from chapter 2. The confidence scores will be tested for their accuracy, with those scores with AUCs above 0.8 being deemed accurate enough to provide utility to the funder. The pre-registration also includes a pre-analysis plan which details hypotheses and subsequent statistical tests regarding the performance of the scores, the characteristics of our participants and original study characteristics amongst others. The pre-registration is split into 3 sections; pre-analysis plan of the ‘meta’ round (the results of which are found in chapter 4), pre-analysis plan of the social and behavioural monthly rounds and the pre-analysis plan of the COVID-19 based round.

Crowdsourced forecasts can also be used to predict scientific outcomes beyond direct replications. The creative destruction approach to replication tests multiple competing theories in the same theoretical space. In this approach, replication simply does not provide supporting or opposing evidence for original findings but rather replaces them with revised theories. I focused on whether forecasters could predict the outcomes of multiple findings relating to gender and hiring decisions (chapter 5) and culture and work ethics (chapter 6). These were the first projects forecasting a creative destruction approach to replication. The findings illustrate that forecasters could predict the outcomes of replications including simple effects, moderator effects and interactions effects and to a lesser extent could also predict the replication effect sizes. However, forecasters were more accurate in predicting the effect sizes relating to culture and work ethics, with a correlation coefficient of 0.7 as opposed to effect sizes relating to gender and hiring decisions with a correlation coefficient of 0.193. I also found that no characteristics of forecasters were correlated with accuracy⁸. These projects were limited as surveys were the only method for eliciting forecasts, where other projects (chapter 2) showed that prediction markets were more accurate.

While the global pandemic created issues for researchers, including myself, by making in-person experiments difficult, it also provided new opportunities for metascience research. Academic publishing witnessed an explosion in research relating to COVID-19 (44), much of which was disseminated on preprint servers (45). I tested whether the future of 400 COVID-19 related preprints could be predicted (chapter 6). We asked forecasters to provide probabilities on 3 publication outcomes, namely (i) published in a high impact journal, (ii) published in a low or

⁸ Correlations here were focused on identifying characteristics of accurate forecasters as opposed to establishing a causal links between a participant's characteristics and their responses.

medium impact journal and (iii) not published within a year. We also elicited forecasts on future citation counts. The sample of 400 preprints was selected by altmetric⁹ score – a measure of impact beyond traditional citations that relies on visibility on social and traditional media. Therefore, we captured a sample with high impact and relevancy. Of the 400 preprints, 69% were published within one year, with 45% of the published papers appearing in high impact journals. On average the preprints received 161 citations, with published preprints receiving more citations than non-published preprints. The results show that forecasts can predict the publication outcomes of the preprints but systematically overestimate the publication rate. Potentially indicating preprints are expected to contain higher quality research than they do. Forecasters additionally proved that the citations of a preprint, ranked relative to the other preprints, can be predicted. Citation rankings were forecasted more accurately than publication outcomes, potentially due to the more stochastic nature of the publication process. Our forecasters were able to predict publication status 58% of the time (correlation of 0.22). This degree of accuracy would need to be improved if this method of assessing preprints was to be implemented in practical uses. Conversely forecasts for citation counts are highly correlated with actual citation counts (correlation of 0.78). While publication and citations are not always perfect proxies for quality, or replicable research (this thesis is testament to this), however it is not unrealistic to expect that peer-review does filter out low quality research to some degree. At the very least, the publication status shows what is deemed to be fit for scientific output and what is not. This research does however have several limitations. Firstly, many of our sample of 400 preprints were already published by the time we undertook forecasts, reducing the sample size. In addition, the preprints had been in circulation for months before forecasting, with many already having citations – providing initial signal to forecasters on final citation ranks. In

⁹ See <https://www.altmetric.com/> for more details

addition, our forecasters were primarily recruited from SCORE – a project focusing on the social and behavioural sciences. Most of the prints were biology and medicine based, and therefore a sample of forecasters who specialised in these areas may have provided more accurate forecasts. This is the first study of its kind and assists in providing indicators of expected quality of non-peer reviewed preprints.

8.2 Future area of study

There are potential avenues for future study that extend the research presented in this thesis.

Firstly, the majority of the projects in this thesis are focused on the social behavioural sciences. There is clear scope to forecast replicability in other fields of science. Issues of reproducibility have been raised in fields such as medicine(46), computer science(11,12), biology(26) and sports science (13). A replicability forecasting project could be conducted in these fields studying the extent to which replicability in these fields can be predicted. Studies selected for forecasting should be systematically sampled across the journals or years to provide unbiased samples such as previous projects (32,33,36). In order to maximise the information gained from such a forecasting project, the number of forecasted claims can be larger than replicated claims – a methodology used in SCORE(41). This allows for inferences of a much larger sample of claims without the resource intensive processes of replication.

Secondly, the methodology used in the projects that make up this thesis can be extended to forecast characteristics of a scientific claim beyond replicability. The characteristics of a claims are multidimensional with aspects such as its direct replicability (as measured in SCORE), how well it generalizes to variation in samples (also known as its generalizability, often measured by conceptual replications), its data analytics replicability (same data, same analysis), its

reproducibility in a multiverse analysis (47,48) where different teams use different analytical approaches to test the same hypotheses, its practicality (how often it will be used both within academia and outside) and the size of the effect described by the claim. While some of these dimensions of a claim may be forecasted in isolation, in the dimensions are all unequivocally related. By exploiting the relationships between these dimensions using combinatorial markets or a Bayesian network analysis (49,50) a forecasting project could forecast multiple dimensions at once.

Thirdly, scientific claims are related or overlap, either supporting or opposing each other. Using the creative destruction approach to replication, these multiple theories are tested at the simultaneously. Current approaches for eliciting assessments of scientific claims treat them in isolation. This approach foregoes an important source of information, namely the judgement of relations between studies. Forecasters who have little information regarding the reliability of two studies may nevertheless be able to make judgements on conditional relations between the studies. It can, for instance, be comparably easy to assess that if one claim is proven is correct, then another cannot be correct. Such conditional relations can be modelled using Bayesian networks and have been shown to improve forecasts on geopolitical events (49,50).

Lastly, the current system of peer-review clearly does not ensure that all published works contain credible findings as 50% of the social and behavioural claims studied in chapter 2 failed to replicate. While you would not expect, nor want, a replication rate of 100%, it is clear that some improvements can be made. Crowd sourced forecasts, through incentivised predictions markets, have been shown to correlate with replicability. There is potential for these forecasts to provide supplemental information to an editor when deciding to accept or reject a paper, either as a 4th reviewer or assisting desk rejections. While forecasters do not provide infallible assessments of credibility, they can provide a valuable assessment as part of the complete peer-review process.

8.3 Final thoughts

The behavioural and social sciences have faced a crisis of trust in the past decade. Researchers who rely on building upon the work of previous established findings, often find that established does not always constitute correct (7,15,51). Large scale replication projects have highlighted shortcomings in both specific findings and field-wide practices. It was not through coincidence that credibility has been diminished – p-hacking, publication bias and other questionable research practices have contributed (1,25,52). In the same way, the road to more credible research will not be shaped by uncoordinated actions. The consequent implementation of open science practices (5) and a focus on correcting past errors can help to restore credibility. Science is not always self -correcting(53), and those that do attempt to correct can often face pushback (54).

We know scientists – and sometimes laypeople - can assess credibility of claims (42,55). This research seeks to develop tools of eliciting information that is widely disseminated amongst the scientific community and aggregating it to useable forecasts. Scientific resources can be sparse, especially when it comes to replications, requiring the use other forms of assessment. Forecasting helps to bridge the gap between highly comprehensive evidence such as that obtained from a many-labs and multiverse analysis, and the lack of strength of evidence in often underpowered original findings. Forecasting can be utilised to assess already published findings to determine expected replicability and using this assessment to allocate replication resource. It can also be used to provide assessments of unpublished papers for peer review or preprints.

While forecasting does not solve the replication crisis, or is 100% accurate, it does provide a valuable addition to the tool belt of credible science.

8.4 References

1. Baker M. 1,500 scientists lift the lid on reproducibility. *Nat News*. 2016 May 26;533(7604):452.
2. Gordon M, Viganola D, Bishop M, Chen Y, Dreber A, Goldfedder B, et al. Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *R Soc Open Sci*. 2020 Jul 22;7(7):200566.
3. Begley CG, Ellis LM. Raise standards for preclinical cancer research. *Nature*. 2012 Mar;483(7391):531–3.
4. Casadevall A, Fang FC. Reproducible Science. *Infect Immun*. 2010 Dec 1;78(12):4972–5.
5. Christensen G, Miguel E. Transparency, Reproducibility, and the Credibility of Economics Research. *J Econ Lit*. 2018 Sep;56(3):920–80.
6. de Waard A, Roffel S, Fennel C, Petridis S, Pijnenburg T, Tsakonas E, et al. Towards Reproducible Artificial Intelligence: Roles and Responsibilities of Researchers and Publishers. 2020;5.
7. Dreber A, Johannesson M. Statistical Significance and the Replication Crisis in the Social Sciences. *Oxf Res Encycl Econ Finance*. 2019 Jul 29 [cited 2020 Jan 17]; Available from: <http://oxfordre.com/view/10.1093/acrefore/9780190625979.001.0001/acrefore-9780190625979-e-461>
8. Drummond DC. Replicability is not Reproducibility: Nor is it Good Science. In 2009 [cited 2020 Feb 8]. Available from: <http://cogprints.org/7691/>
9. Fanelli D. Negative results are disappearing from most disciplines and countries. *Scientometrics*. 2012 Mar 1;90(3):891–904.
10. Fanelli D. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proc Natl Acad Sci*. 2018 Mar 13;115(11):2628–31.
11. Gundersen OE, Gil Y, Aha DW. On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. *AI Mag*. 2018 Sep 28;39(3):56–68.
12. Gundersen OE, Kjensmo S. State of the Art: Reproducibility in Artificial Intelligence. In: *Thirty-Second AAAI Conference on Artificial Intelligence 2018* [cited 2020 Feb 19]. Available from: <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17248>
13. Halperin I, Vigotsky AD, Foster C, Pyne D. Strengthening the Practice of Exercise and Sport-Science Research. *Int J Sports Physiol Perform*. 2018;

14. Hunter JE. The Desperate Need for Replications. *J Consum Res.* 2001 Jun 1;28(1):149–58.
15. Ioannidis J. Why Most Published Research Findings Are False. *PLOS Med.* 2005 Aug 30;2(8):e124.
16. Ioannidis J, Doucouliagos C. What’s to Know About the Credibility of Empirical Economics? *J Econ Surv.* 2013;27(5):997–1004.
17. McCullough B d., McGeary KA, Harrison TD. Do economics journal archives promote replicable research? *Can J Econ Can Déconomique.* 2008;41(4):1406–20.
18. Millstone E, van Zwanenberg P. A crisis of trust: for science, scientists or for institutions? *Nat Med.* 2000 Dec;6(12):1307–8.
19. Mobley A, Linder SK, Braeuer R, Ellis LM, Zwelling L. A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic. *PLoS ONE* 2013 May 15 [cited 2021 Mar 24];8(5). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3655010/>
20. Morrison SJ. Time to do something about reproducibility. *eLife.* 2014 Dec 10;3:e03981.
21. Nosek BA, Hardwicke TE, Moshontz H, Allard A, Corker KS, Almenberg AD, et al. Replicability, Robustness, and Reproducibility in Psychological Science. *PsyArXiv*; 2021 [cited 2021 Mar 25]. Available from: <https://psyarxiv.com/ksfvq/>
22. Pashler H, Wagenmakers E. Editors’ Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspect Psychol Sci.* 2012 Nov 1;7(6):528–30.
23. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov.* 2011 Sep;10(9):712–712.
24. Schloss PD. Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. *mBio.* 2018 Jul 5 [cited 2021 Apr 14];9(3). Available from: <https://mbio-asm-org.ezproxy.massey.ac.nz/content/9/3/e00525-18>
25. Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol Sci.* 2011 Nov 1;22(11):1359–66.
26. Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA. An open investigation of the reproducibility of cancer biology research. *Rodgers P, editor. eLife.* 2014 Dec 10;3:e04333.

27. Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, Banks JB, et al. Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J Exp Soc Psychol*. 2016 Nov 1;67:68–82.
28. Ebersole CR, Mathur MB, Baranski E, Bart-Plange D-J, Buttrick NR, Chartier CR, et al. Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. *Adv Methods Pract Psychol Sci*. 2020 Sep 1;3(3):309–31.
29. Klein RA, Ratliff KA, Vianello M, Adams Jr. RB, Bahník Š, Bernstein MJ, et al. Investigating variation in replicability: A “many labs” replication project. *Soc Psychol*. 2014;45(3):142–52.
30. Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB, Alper S, et al. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Adv Methods Pract Psychol Sci*. 2018 Dec;1(4):443–90.
31. Klein RA, Cook CL, Ebersole CR, Vitiello C, Nosek BA, Chartier CR, et al. Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement *PsyArXiv*; 2019 [cited 2021 Mar 24]. Available from: <https://psyarxiv.com/vef2c/>
32. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015 Aug 28;349(6251):aac4716.
33. Camerer CF, Dreber A, Forsell E, Ho T-H, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. *Science*. 2016 Mar 25;351(6280):1433–6.
34. Chang AC, Li P. Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say “Usually Not” Rochester, NY: Social Science Research Network; 2015 Sep [cited 2021 Sep 23]. Report No.: ID 2669564. Available from: <https://papers.ssrn.com/abstract=2669564>
35. Cova F, Strickland B, Abatista A, Allard A, Andow J, Attie M, et al. Estimating the Reproducibility of Experimental Philosophy. *Rev Philos Psychol*. 2021 Mar 1;12(1):9–44.
36. Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nat Hum Behav*. 2018 Sep;2(9):637–44.
37. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. *Nat Hum Behav*. 2018 Jan;2(1):6.
38. Atanasov P, Rescober P, Stone E, Swift SA, Servan-Schreiber E, Tetlock P, et al. Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls. *Manag Sci*. 2017 Mar;63(3):691–706.

39. Wang G, Kulkarni SR, Poor HV, Osherson DN. Aggregating Large Sets of Probabilistic Forecasts by Weighted Coherent Adjustment. *Decis Anal*. 2011 Jun 1;8(2):128–44.
40. Baron J, Mellers BA, Tetlock PE, Stone E, Ungar LH. Two Reasons to Make Aggregated Probability Forecasts More Extreme. *Decis Anal* 2014 Mar 19 [cited 2019 Mar 20]; Available from: <https://pubsonline.informs.org/doi/abs/10.1287/deca.2014.0293>
41. Alipourfard N, Arendt B, Benjamin DM, Benkler N, Bishop M, Burstein M, et al. Systematizing Confidence in Open Research and Evidence (SCORE) SocArXiv; 2021 [cited 2021 Jun 14]. Available from: <https://osf.io/preprints/socarxiv/46mnb/>
42. Gordon M, Viganola D, Dreber A, Johannesson M, Pfeiffer T. Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. *PLOS ONE*. 2021 Apr 14;16(4):e0248780.
43. Liu Y, Wang J, Chen Y. Surrogate Scoring Rules. *ACM Conf Econ Comput* 2020 Jul; Available from: <http://arxiv.org/abs/1802.09158>
44. Nowakowska J, Sobocińska J, Lewicki M, Lemańska Ż, Rzymiski P. When science goes viral: The research response during three months of the COVID-19 outbreak. *Biomed Pharmacother*. 2020 Sep 1;129:110451.
45. Älgå A, Eriksson O, Nordberg M. The development of preprints during the COVID-19 pandemic. *J Intern Med*. 2021;290(2):480–3.
46. Ioannidis JPA. Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *JAMA*. 2005 Jul 13;294(2):218–28.
47. Klapwijk ET, van den Bos W, Tamnes CK, Raschle NM, Mills KL. Opportunities for increased reproducibility and replicability of developmental neuroimaging. *Dev Cogn Neurosci*. 2021 Feb 1;47:100902.
48. Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*. 2020 Jun;582(7810):84–8.
49. Laskey KB, Hanson R, Twardy C. Combinatorial prediction markets for fusing information from distributed experts and models. In: 2015 18th International Conference on Information Fusion (Fusion). 2015. p. 1892–8.
50. Sun W, Hanson R, Laskey KB, Twardy C. Probability and Asset Updating using Bayesian Networks for Combinatorial Prediction Markets. 2012 Oct 16 [cited 2020 Mar 9]; Available from: <https://arxiv.org/abs/1210.4900v1>
51. Vazire S. Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspect Psychol Sci*. 2018 Jul 1;13(4):411–7.

52. John LK, Loewenstein G, Prelec D. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychol Sci*. 2012 May 1;23(5):524–32.
53. Ioannidis JPA. Why Science Is Not Necessarily Self-Correcting. *Perspect Psychol Sci*. 2012 Nov 1;7(6):645–54.
54. Vazire S. A toast to the error detectors. *Nature*. 2019 Dec 30;577(7788):9–9.
55. Hoogeveen S, Sarafoglou A, Wagenmakers E-J. Laypeople Can Predict Which Social-Science Studies Will Be Replicated Successfully. *Adv Methods Pract Psychol Sci*. 2020 Sep 1;3(3):267–85.

9 Appendices

9.1 Appendix 1: Systematizing Confidence in Open Research and Evidence (SCORE)

This appendix contains the paper “Systematizing Confidence in Open Research and Evidence (SCORE)” that is currently under review at “Proceedings of the National Academy of Sciences of the United States of America”. This paper has also been released as a preprint. This paper describes the SCORE project in its entirety, including the contributions of each team. The team I was part of was the ‘Replication Markets’. This paper has a large scope, much of it beyond the research I was directly involved in, however it is included in this thesis as an appendix it provides a complete and detailed overview of the SCORE project, providing context to interpret any research relevant to SCORE. My contribution to this paper is limited to drafting of replication markets section

This chapter is currently under review at the journal ‘Proceedings of the National Academy of Sciences of the United States of America’. It is also uploaded on the preprint server SocArXiv. The reference for the print is: Alipourfard N, Arendt B, Benjamin DJ, Benkler N, Bishop M, Burstein M, Bush M, Caverlee J, Chen Y, Clark C, Almenberg AD. Systematizing Confidence in Open Research and Evidence (SCORE).

To align with the formatting and referencing style of this thesis, there are some changes in formatting and referencing style of the preprint

Systematizing Confidence in Open Research and Evidence (SCORE)

9.1.1 Abstract

Assessing the credibility of research claims is a central, continuous, and laborious part of the scientific process. Credibility assessment strategies range from expert judgment to aggregating existing evidence to systematic replication efforts. Such assessments can require substantial time and effort. Research progress could be accelerated if there were rapid, scalable, accurate credibility indicators to guide attention and resource allocation for further assessment. The SCORE program is creating and validating algorithms to provide confidence scores for research claims at scale. To investigate the viability of scalable tools, teams are creating: a database of claims from papers in the social and behavioral sciences; expert and machine generated estimates of credibility; and, evidence of reproducibility, robustness, and replicability to validate the estimates. Beyond the primary research objective, the data and artifacts generated from this program will be openly shared and provide an unprecedented opportunity to examine research credibility and evidence.

147 Words

Keywords: Metascience, replicability, reproducibility, social sciences, credibility, algorithms

¹ Co-authors listed alphabetically: Nazanin Alipourfard , University of Southern California ; Beatrix Arendt , Center for Open Science ; Daniel Benjamin , University of Southern California ; Noam Benkler , SIFT ; Mark Burstein , SIFT ; Martin Bush , University of Melbourne ; James Caverlee , Texas A&M University ; Yiling Chen , Harvard University ; Chae Clark , TwoSix Technologies ; Anna Dreber , Stockholm School of Economics ; Timothy M. Errington , Center for Open Science ; Fiona Fidler , University of Melbourne ; Nicholas Fox , Center for Open Science ; Aaron Frank , RAND Corporation ; Hannah Fraser , University of Melbourne ; Scott Friedman , SIFT ; Ben Gelman , TwoSix Technologies ; James Gentile , TwoSix Technologies ; C Lee Giles , The Pennsylvania State University ; Michael Gordon , Massey University; Reed Gordon-Sarney , TwoSix Technologies ; Christopher Griffin , The Pennsylvania State University ; Timothy Gulden , RAND Corporation ; Krystal Hahn , Center for Open Science ; Robert Hartman , The MITRE Corporation ; Felix Holzmeister , University of Innsbruck ; Xia Hu , Texas A&M University ; Magnus Johannesson , Stockholm School of Economics ; Lee Kezar , University

of Southern California ; Melissa Kline Struhl , Center for Open Science ; Ugur Kuter , SIFT ; Anthony Kwasnica , The Pennsylvania State University ; Dong-Ho Lee

, University of Southern California ; Kristina Lerman , University of Southern California ; Yang Liu , University of California, Santa Cruz ; Zach Loomas , Center for Open Science ; Bri Luis , Center for Open Science ; Ian Magnusson , SIFT ; Michael Bishop , Ottawa, ON, Canada ; Olivia Miske , Center for Open Science ; Fallon Mody , University of Melbourne ; Fred Morstatter , University of Southern California ; Brian A. Nosek , Center for Open Science; University of Virginia ; E. Simon Parsons , Center for Open Science ; David Pennock , Rutgers University ; Thomas Pfeiffer , Massey University; Haochen Pi , University of Southern California ; Jay Pujara , University of Southern California ; Sarah Rajtmajer , The Pennsylvania State University ; Xiang Ren , University of Southern California ; Abel Salinas , University of Southern California ; Ravi Selvam , University of Southern California ; Frank Shipman , Texas A&M University ; Priya Silverstein , Center for Open Science; Institute for Globally Distributed Open Research and Education ; Amber Sprenger , The MITRE Corporation ; Anna Squicciarini , The Pennsylvania State University ; Stephen Stratman , The MITRE Corporation ; Kexuan Sun , University of Southern California ; Saatvik Tikoo , University of Southern California ; Charles R. Twardy , Jacobs / George Mason ; Andrew Tyner , Center for Open Science ; Domenico Viganola , World Bank ; Juntao Wang , Harvard University ; David Wilkinson , University of Melbourne ; Bonnie Wintle , University of Melbourne ; Jian Wu , Old Dominion University

9.1.2 Authors' note

This work was supported by the Defense Advanced Research Projects Agency. This paper is authored by members of the teams that are directly involved in the SCORE program, but many of the activities - both those funded by the SCORE program itself and the other scientific collaborations that are beginning to form - involve both extensive staff within each team and formal and informal collaborations at other institutions. While this full network of contributors are not authors on this paper, they are critical to the program's execution. Future articles and other scientific contributions resulting from SCORE will be carried out both by large collaborative teams and by the smaller lists of authors that are more typical for many of the fields represented in this program. Co-authors with an affiliation to the Center for Open Science acknowledge a conflict of interest as employees of the nonprofit organization with a mission to increase openness, integrity, and reproducibility of research.

9.1.3 Authors' contributions

For this paper, contributions are summarized using the following categories:

- A. Contributed descriptions of their organizations' specific roles in the program
- B. Contributed significantly to the writing of other sections (e.g. introduction and conclusion).
- C. Integrated these sections to construct the first draft.
- D. Provided feedback and revisions to produce the submitted version of this manuscript.
- E. Drafted initial structure and outline of the paper

Given Name	Family Name	Institution(s)	Location	Contribution
Nazanin	Alipourfard	University of Southern California	Los Angeles, CA	A
Beatrix	Arendt	Center for Open Science	Charlottesville, VA USA	D
Daniel	Benjamin	University of Southern California	Los Angeles, CA	A
Noam	Benkler	SIFT	Minneapolis, MN, USA	A
Michael	Bishop		Ottawa, ON, CANADA	A
Mark	Burstein	SIFT	Minneapolis, MN, USA	A
Martin	Bush	University of Melbourne	Melbourne, Australia	A, D
James	Caverlee	Texas A&M University	College Station, TX, USA	A
Yiling	Chen	Harvard University	Cambridge, MA USA	A
Chae	Clark	TwoSix Technologies	Arlington, VA, USA	A
Anna	Dreber	Stockholm School of Economics	Stockholm, SWEDEN	A
Timothy M.	Errington	Center for Open Science	Charlottesville, VA, USA	A, C, D, E
Fiona	Fidler	University of Melbourne	Melbourne, Australia	A

Nicholas	Fox	Center for Open Science	Charlottesville, VA USA	A, D
Aaron	Frank	RAND Corporation	Arlington, VA	A

Hannah	Fraser	University of Melbourne	Melbourne, Australia	A
Scott	Friedman	SIFT	Minneapolis, MN, USA	A
Ben	Gelman	TwoSix Technologies	Arlington, VA, USA	A
James	Gentile	TwoSix Technologies	Arlington, VA, USA	A
C Lee	Giles	The Pennsylvania State University	State College, PA, USA	A
Michael	Gordon	Massey University	Auckland, NEW ZEALAND	A
Reed	Gordon-Sarney	TwoSix Technologies	Arlington, VA, USA	A
Christopher	Griffin	The Pennsylvania State University	State College, PA, USA	A
Timothy	Gulden	RAND Corporation	Santa Monica, CA	A
Krystal	Hahn	Center for Open Science	Charlottesville, VA USA	D
Robert	Hartman	The MITRE Corporation	McLean, VA	A
Felix	Holzmeister	University of Innsbruck	Innsbruck, AUSTRIA	A
Xia	Hu	Texas A&M University	College Station, TX, USA	A
Magnus	Johannesson	Stockholm School of Economics	Stockholm, SWEDEN	A
Lee	Kezar	University of Southern California	Los Angeles, CA	A
Melissa	Kline Struhl	Center for Open Science	Charlottesville, VA, USA	A, B, C, D, E
Ugur	Kuter	SIFT	Minneapolis, MN, USA	A
Anthony	Kwasnica	The Pennsylvania State University	State College, PA, USA	A
Dong-Ho	Lee	University of Southern California	Los Angeles, CA	A
Kristina	Lerman	University of Southern California	Los Angeles, CA	A
Yang	Liu	University of California, Santa Cruz	Santa Cruz, CA, USA	A
Zach	Loomas	Center for Open Science	Charlottesville, VA USA	C, D
Bri	Luis	Center for Open Science	Charlottesville, VA USA	D
Ian	Magnusson	SIFT	Minneapolis, MN, USA	A
Olivia	Miske	Center for Open Science	Charlottesville, VA USA	C, D
Fallon	Mody	University of Melbourne	Melbourne, Australia	A
Fred	Morstatter	University of Southern California	Los Angeles, CA	A
Brian A.	Nosek	Center for Open Science; University of Virginia	Charlottesville, VA USA	A, B, C, D, E
E. Simon	Parsons	Center for Open Science	Charlottesville, VA USA	D
David	Pennock	Rutgers University	New Brunswick, NJ, USA	A
Thomas	Pfeiffer	Massey University	Auckland, NEW ZEALAND	A
Haochen	Pi	University of Southern California	Los Angeles, CA	A
Jay	Pujara	University of Southern California	Los Angeles, CA	A
Sarah	Rajtmajer	The Pennsylvania State	State College, PA, USA	A

		University		
Xiang	Ren	University of Southern California	Los Angeles, CA	A
Abel	Salinas	University of Southern California	Los Angeles, CA	A
Ravi	Selvam	University of Southern California	Los Angeles, CA	A
Frank	Shipman	Texas A&M University	College Station, TX, USA	A
Priya	Silverstein	Center for Open Science; Institute for Globally Distributed Open Research and Education	Charlottesville, VA, USA; Preston, UK	C, D
Amber	Sprenger	The MITRE Corporation	McLean, VA	A
Anna	Squicciarini	The Pennsylvania State University	State College, PA, USA	A
Stephen	Stratman	The MITRE Corporation	McLean, VA	A
Kexuan	Sun	University of Southern California	Los Angeles, CA	A
Saatvik	Tikoo	University of Southern California	Los Angeles, CA	A
Charles R.	Twardy	Jacobs / George Mason	Herndon/Fairfax, VA USA	A
Andrew	Tyner	Center for Open Science	Charlottesville, VA USA	A, B
Domenico	Viganola	World Bank	Washington, DC, USA	A
Juntao	Wang	Harvard University	Cambridge, MA, USA	A
David	Wilkinson	University of Melbourne	Melbourne, Australia	A
Bonnie	Wintle	University of Melbourne	Melbourne, Australia	A
Jian	Wu	Old Dominion University	Norfolk, VA, USA	A

A primary activity of science is evaluating the credibility of claims--assertions reported as findings from the evaluation of evidence. Researchers create evidence and make claims about what that evidence means. Others assess those claims to determine their credibility including assessing reliability, validity, generalizability, and applicability. Assessment occurs by journal reviewers during the peer review process; by readers deciding whether claims should inform their judgment; by researchers trying to replicate, extend, confirm, or challenge prior claims; by funders deciding what is worth further investment; and by practitioners and policymakers determining whether the claims should inform policy or practice.

Assessing confidence in research claims is important and resource intensive. A reader must read and think about a paper to assess confidence in its claims against their expert judgment and reasoning. A researcher expends substantial effort planning, conducting, and reporting follow up

research to assess the credibility of prior claims. Rarely is a single follow up investigation the end of the story. Researchers may go back and forth for multiple years challenging, debating, and refining their understanding of claims. And, sometimes it is difficult or impossible to obtain additional evidence; A decision needs to be made about credibility with only what is already available.

The “Systematizing Confidence in Open Research and Evidence” (SCORE) program has an aspirational objective to develop and validate methods to assess the credibility of research claims at scale with much greater speed and much lower cost than is possible at present. Imagine it takes a year to achieve 95% accuracy in assessing the credibility of a claim by conducting replication and generalizability studies, a month to achieve 85% accuracy by conducting reproduction and robustness tests of the same claim, and a few hours to achieve 80% accuracy by consulting a group of experts to review the readily available evidence. Could we create automated methods to achieve similar accuracy as experts in a few minutes or a few seconds? If that were possible, readers, researchers, reviewers, funders, and policymakers could use the rapid assessments to direct their attention for more laborious assessment and improve judicious allocation of resources to examine claims that are important but relatively uncertain or low in confidence.

There is accumulating evidence that such a service is needed and possible to achieve. In the social and behavioral sciences, replication efforts have indicated that the literature is not as replicable as might be expected (1–8). For example, Nosek and colleagues (9) aggregated 307 replication attempts of published findings in psychology and observed that 64% reported statistically significant evidence in the same direction as the original studies, with effect sizes 68% as large as the original studies. Investigations of robustness and reproducibility of claims suggest that some published evidence is highly contingent on specific analytic decisions, or even

irreproducible(10–12). These investigations indicate that the credibility of published claims is more uncertain than expected.

Multiple studies indicate that people can anticipate which findings are likely to replicate after reading the original paper or even just reviewing a subset of information about the finding and supporting evidence (1,2,13–15). Human judgments were correlated with successful replication using prediction markets ($r = 0.52$), surveys ($r = 0.48$), and structured elicitations ($r = 0.75$; see Nosek et al.(9) for a review). This provides initial evidence that relatively accurate credibility assessments are achievable with an order (or orders) of magnitude lower resource investment than conducting replication or reproduction studies.

Finally, three studies provide initial evidence that machine learning methods may provide a scalable solution that could match, or perhaps even exceed, the capabilities of human judgment(16–18). Each machine learning investigation used a distinct approach drawing on narrative text of the original paper, information about original designs and replication sample sizes, or other contextual information about the original finding. These promising findings provide a basis for SCORE’s primary goal to investigate scalable methods of assessing credibility of claims in the social-behavioral sciences.

SCORE began in February 2019 and the main activities are expected to conclude in May 2022. This paper introduces the program structure, activities, and expected outcomes of the program, including data and artifacts that will be made available to the research community for further investigation.

9.1.4 Program Scope and Structure

SCORE is a large-scale collaboration involving eight primary research teams and more than a thousand contributing researchers. The teams are organized into three technical areas (TAs) - TA1, TA2, and TA3 - and a Testing and Evaluation (T&E) group that evaluates the TAs and program effectiveness. The primary research teams have clearly specified roles, distinct areas of expertise, and shared objectives organized around a common set of articles constituting the shared Common Task Framework (CTF). The research teams work with the shared CTF dataset in a coordinated way to advance the SCORE program's goals (see Figure 1).

The CTF consists of approximately 30,000 articles from 2009-2018, representing 62 journals from the following disciplines: Criminology, Economics and Finance, Education, Health, Management, Marketing and Organizational Behavior, Political Science, Psychology, Public Administration, and Sociology (see Table 1). From the CTF, a stratified random sample of 3,000 papers was selected for additional investigation and enhancement, called the *annotation set*. From the annotation set, a stratified random sample of 600 papers was then sampled for additional investigation such as conducting reproduction or replication studies, called the *evidence set*. This sampling was done without regard to the feasibility of any particular empirical attempt, with the understanding that not all claims will receive a completed empirical study result. This design is intended to be adaptive to the resource-intensiveness of different activities for assessing credibility while also maximizing the generalizability of the findings to the social-behavioral sciences.

Table 9-1. Journals comprising the Common Task Framework (CTF)

Discipline	Journals
Criminology	<i>Law and Human Behavior</i> <i>Criminology</i>

Economics and Finance	<i>Experimental Economics Journal</i> <i>of Labor Economics</i> <i>The Quarterly Journal of Economics</i> <i>Journal of Political Economy</i> <i>Econometrica</i> <i>American Economic Review</i> <i>The Journal of Finance</i> <i>Journal of Financial Economics</i> <i>American Economic Journal: Applied Economics</i> <i>Review of Financial Studies</i>
Education	<i>American Educational Research Journal</i> <i>Exceptional Children</i> <i>Computers & Education</i> <i>Contemporary Educational Psychology</i> <i>Educational Researcher</i> <i>Journal of Educational Psychology</i> <i>Learning and Instruction</i>
Health	<i>Psychological Medicine</i> <i>Health Psychology</i> <i>Social Science & Medicine</i>
Management	<i>Journal of Business Research</i> <i>The Leadership Quarterly</i> <i>Academy of Management Journal</i> <i>Management Science</i> <i>Journal of Management</i> <i>Organization Science</i>
Marketing and Organizational Behavior	<i>Journal of Consumer Research</i> <i>Journal of the Academy of Marketing Science</i> <i>Journal of Organizational Behavior</i> <i>Journal of Marketing</i> <i>Journal of Marketing Research</i> <i>Organizational Behavior and Human Decision Processes</i>
Political Science	<i>Journal of Experimental Political Science</i> <i>American Journal of Political Science</i> <i>American Political Science Review</i> <i>World Politics</i> <i>British Journal of Political Science</i>

	<i>Journal of Conflict Resolution</i> <i>Comparative Political Studies</i> <i>World Development</i>
Psychology	<i>Journal of Experimental Social Psychology</i> <i>Journal of Applied Psychology</i> <i>Journal of Environmental Psychology</i> <i>Journal of Personality and Social Psychology</i> <i>Journal of Experimental Psychology: General</i> <i>Evolution and Human Behavior</i> <i>Psychological Science</i> <i>Cognition</i> <i>European Journal of Personality</i> <i>Child Development</i> <i>Journal of Consulting and Clinical Psychology</i> <i>Clinical Psychological Science</i>
Public Administration	<i>Journal of Public Administration Research and Theory</i> <i>Public Administration Review</i>
Sociology	<i>Journal of Marriage and Family</i> <i>American Sociological Review</i> <i>American Journal of Sociology</i> <i>Demography</i> <i>Social Forces</i> <i>European Sociological Review</i>

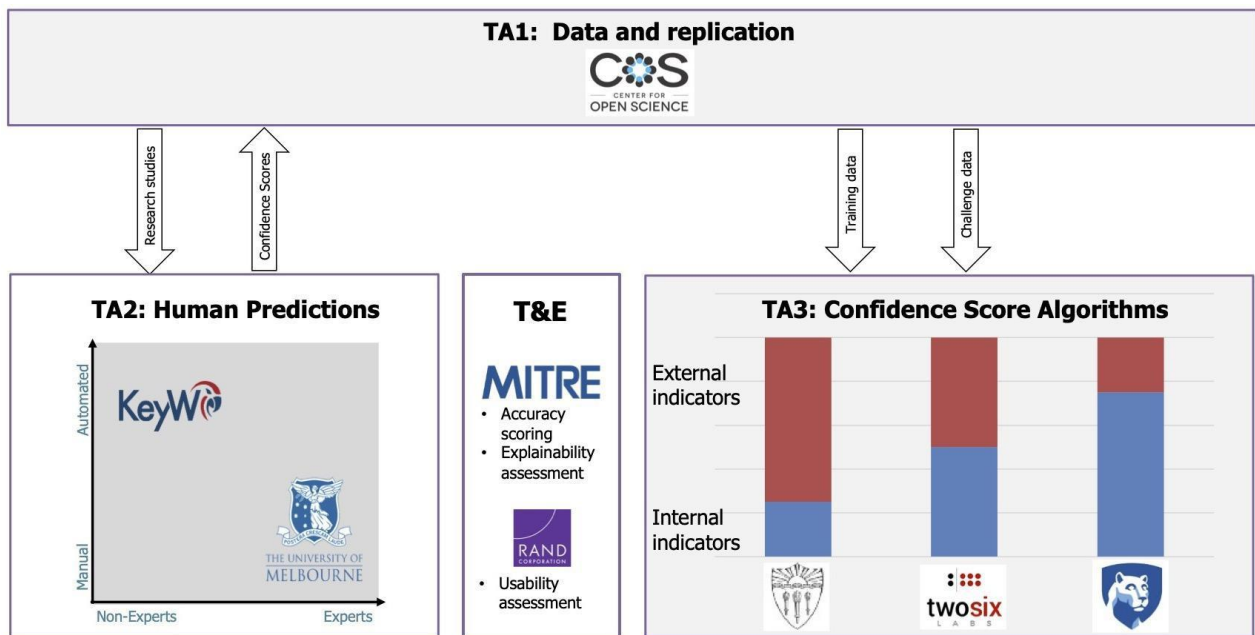
The purpose of the team structure and shared set of papers is to investigate the credibility of claims from the social-behavioral sciences and test methods for efficiently assessing that credibility. To do this, the project is organized in modular stages with specific responsibilities for each team.

TA1, the Center for Open Science (COS), is responsible for enhancing the CTF database and extracting claims for the *annotation set* to be evaluated by the other teams. In the completed half of the program, this meant extracting 3,000 individual claims, one from each paper. TA2 teams from KeyW/Jacobs Corporation and University of Melbourne used human evaluators to provide confidence scores predicting the reproducibility or replicability of the 3,000 research claims in the *annotation set*. These teams competed with each other to provide the most accurate scores. Three TA3 teams

from Pennsylvania State University (PSU), TwoSix Labs, LLC, and University of Southern California (USC) used machine learning methods to develop algorithms that assign confidence scores just like the human evaluators.

While TA2 and TA3 teams generated scores for these 3,000 claims, TA1 privately created a stratified random sample of 600 of those papers to create the *evidence set*. Some claims from the evidence set were subjected to reproduction and replication studies. TA2 and TA3 teams were left unaware of which claims were selected for the *evidence set* to avoid any complications of altering strategy to focus on specific claims. The reproduction and replication outcomes provide a ground truth benchmark for evaluating accuracy of the confidence scores generated by humans, a process managed by the Testing and Evaluation (T&E) teams. Algorithms are evaluated primarily on their ability to predict the human credibility assessments across all 3,000 claims, and assessed for explainability of the generated confidence scores. Which claims were selected for replication or reproduction studies, and the outcomes of those studies are held back from TA2 and TA3 teams until their credibility scores are committed and completed.

Figure 9-1. Relationships between research teams comprising the three technical areas (TAs) of the SCORE program.



Entering the second half of the program, the breadth and depth of the project is expanding with TA1 sampling additional claims from the CTF, extracting a single claim per paper for another 900 papers, and systematically extracting a complete “bushel” of claims from 200 of the initial 600 papers in the *evidence set*. The complete set of bushel claims is meant to represent all of the claims that could have been selected from the paper in the first half of the program, rather than simply the one claim that was selected. The Melbourne TA2 team is expanding the task of the human evaluators to evaluate all of the bushel claims and to assess those papers on multiple indicators of credibility. TA3 teams are extending their strategies for improving algorithm performance. And, finally, TA1 is expanding the scope of assessing reproduction, robustness, and replicability for the *evidence set* of 600 papers.

9.1.5 What Makes SCORE Unique

SCORE draws inspiration from prior research on systematic replications and reproductions(1–8,19–22) and replicability predictions by humans (2,13,14) and

machines (16–18). SCORE extends these efforts in both its unprecedented scale and its disciplinary scope. The sampling strategy is inclusive of a substantial portion of the social-behavioral sciences to facilitate generalizability and investigation of heterogeneity in credibility and replicability across subdisciplines and methodologies. Also, with a standard identification process of discrete claims across papers, the SCORE program facilitates broad inclusion of outcome types, comparison of those outcomes across papers, and a variety of verification attempts including reproduction, robustness, and replication tests.

Another virtue of the SCORE program is that it includes many distinct efforts on the same large dataset, facilitating the opportunity for comparative analysis. For example, the most enriched papers from the *evidence set* will have structured claim extraction from the paper, metadata about the paper from external databases (e.g., citation rates, presence of open data), human credibility scores from multiple sources, machine credibility scores from multiple sources, and evidence on reproducibility, robustness, and replicability of one or multiple claims. This accumulated data will facilitate many investigations beyond the primary objective of SCORE.

Finally, like prior large-scale replication projects, at the conclusion of the program, SCORE data will be accessible to others for research. Additional users of SCORE data may themselves enhance the dataset and other artifacts creating a generative, virtuous cycle of data enrichment fostering new investigations that provide further enrichment.

9.1.6 Defining and Extracting Scientific Claims

The TA1 team is responsible for annotating the papers randomly sampled into the *annotation set*. In the completed first half of the project, this meant identifying a single relevant *claim* from each paper, by tagging related information from the pdf of an article. In SCORE terminology, this claim represents a specific, concrete finding that is supported by a statistically significant test result, or at least by evidence that would be amenable to a statistical hypothesis test even if the authors did not adopt significance testing. This is not the only way to identify a claim, but this working definition provides clarity between teams, sufficient flexibility to cover a wide range of research applications, and is sufficient constraint to define criteria for evaluating confidence and assessing replicability and reproducibility. Table 2 shows a glossary of working definitions used in SCORE.

Table 9-2: A glossary of key terms as they are used for the SCORE program

Paper	A single academic article that makes quantitative claims based on specific social scientific data. SCORE does not address papers that are exclusively based on qualitative research, simulations, theory, or commentary.
Common Task Framework (CTF)	The set of approximately 30,000 papers that constitutes the sampling frame for SCORE. It includes papers from 62 social science journals published between 2009 and 2018.
Annotation Set	A stratified random sample from the CTF of approximately 3,000 papers that are annotated to identify at least one claim trace per paper.
Evidence Set	A stratified random sample of approximately 600 papers from the annotation set. These papers could be selected for an empirical attempt to find further evidence for or against a claim.
Claim	A specific assertion reported as a finding in a paper. Most papers make more than one claim, and claims in a paper can be related or independent of one another.
Claim Trace	A claim in a paper is identified by annotating and labeling short excerpts from the main text or tables/graphs from the paper. Together these annotations let a reader ‘trace’ from a general statement in the abstract to a more specific claim to the quantitative information such as a specific inferential test or estimate that is given as evidence for that claim.

Confidence Score	A prediction about the replicability of a claim, expressed as a numerical value on a scale from “not confident” to “very confident.” Confidence scores are about a single claim which may or may not generalize to confidence in other claims from the same paper.
Inferential Test	A statistical calculation that supports an inference about a single effect and provides information about both the spread and central tendency of that effect. When testing statistical significance, a single inferential test is associated with a single p value. Additionally, with regression modeling, inferential tests may be associated with a single parameter, or with an entire model if model comparison tests are conducted.
Bushel Claim	A set of claim traces from a single paper representing as many of the independent claim traces that the authors present as possible. Each claim trace must be linked to a finding reported in the abstract, and must be supported by quantitative evidence presented in the main text.
Empirical Study	A single empirical attempt conducted by a research team to provide additional evidence about a claim. These attempts can include conducting a replication, reproduction, or other empirical activity that speaks to the credibility of that claim.
Replication	Testing the reliability of a prior finding with new data expected to be theoretically equivalent by comparing the outcome of an inferential test as reported in a paper with the equivalent inferential test as calculated in the new dataset.
Reproduction	Testing the reliability of a prior finding with the same data and same analysis strategy by comparing the outcome of an inferential test as reported in a paper with a re-calculation of that inferential test from the original data.
Robustness	Testing the reliability of a prior finding with the same data and different analysis strategy by conducting alternative tests on the original data.
Generalizability	Testing the reliability of a prior finding in a new dataset in a way that differs from the original study but is expected to produce similar results.

Table 9-3: A single claim trace of a paper is composed of four levels.

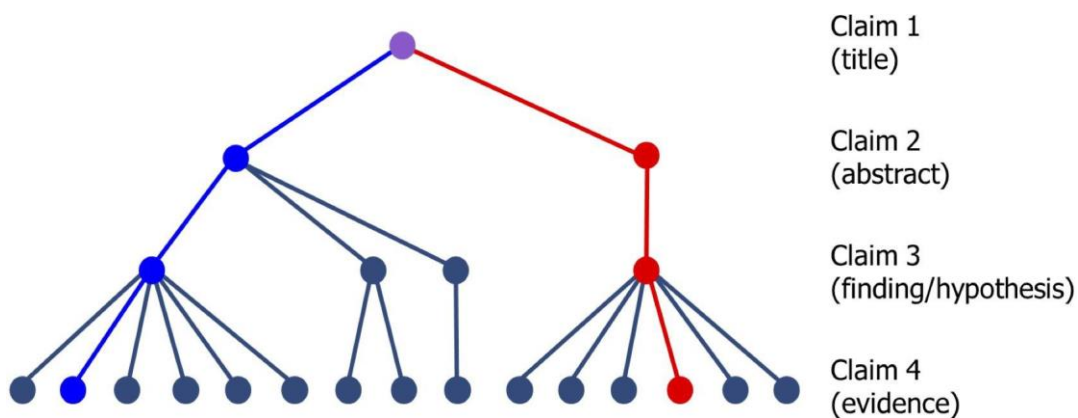
Claim 1	The title of the paper--the most general statement of a topic or finding.
Claim 2	A statement from the paper’s abstract that reflects an empirical research finding.
Claim 3	A hypothesis, prediction, or finding statement presented somewhere in the main text of the paper, relating to the finding reported in Claim 2.
Claim 4	A result supported by specific statistical information in the article that supports Claim 3, alongside the authors’ interpretation of that information.

The output of the annotation process is a “claim trace” that maps a finding reported in the abstract to a specific hypothesis or finding statement from the main text, to a particular set of quantitative evidence that supports the reported finding. When only

one claim trace is identified in a paper, the process does not guarantee that the claim trace selected necessarily includes the paper’s “most important” or “most central” claim. This kind of decision is neither objective nor obvious for many papers. Pretesting revealed that such a standard is difficult to define.

Instead, as a proxy for a lower bound on importance, a claim must be directly related to a statement made in the paper’s abstract. This criterion avoids selecting tangential findings that are not related to the summarized purpose of the paper. The claim trace indicates a series of levels leading down to the specific focal result as described in Table 3.

Figure 9-2. Model of a bushel claim set for a single paper. Each line represents a distinct bushel claim trace. Two examples of single-trace claims that could have been extracted are in blue and red.



Selecting a single finding creates a tractable and comparable way for independent teams to work with a paper, and it has clear limitations for interpreting the results. Papers often include more than one finding in the abstract, and research findings are often supported by multiple pieces of evidence. In the current phase of work, we have expanded claim extraction for some papers in the *evidence set* by adding a second bushel approach that relaxes these requirements. In the bushel approach, we identify as many unique claims as possible by tracing from a finding in the abstract to statistical

evidence in the paper. In addition, we relax the definitions of evidence to allow tagging of multiple inferential tests and other types of quantitative evidence. Figure 2 illustrates a bushel of claims from a paper and two single-trace claims that could be extracted.

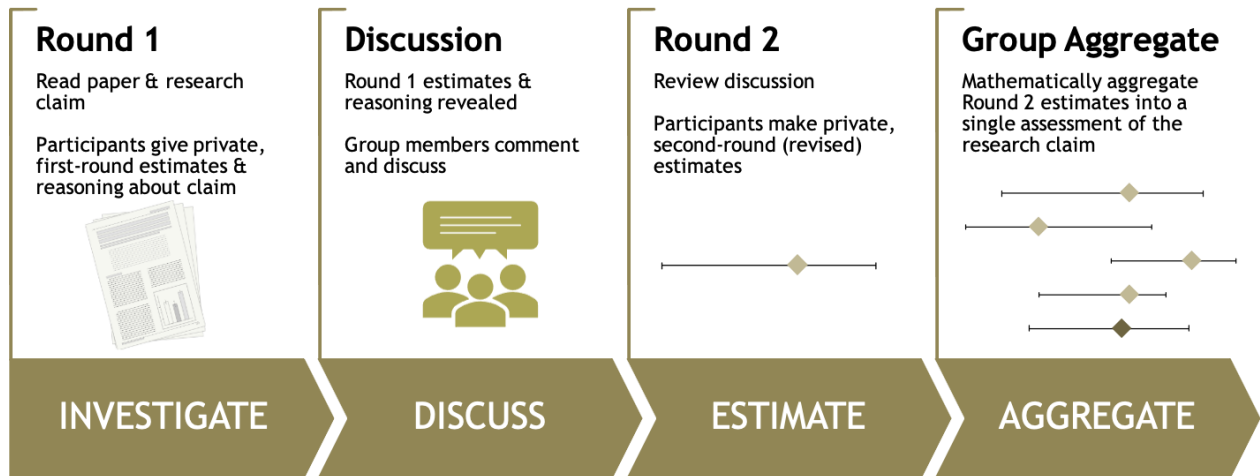
9.1.7 Expert Assessment

The second major technical area (TA2) elicits predictions, called confidence scores, from human readers about replicability of extracted claims. TA2 included two independent teams, repliCATS and Replication Markets, to examine the viability and accuracy of distinct forecasting strategies.

9.1.7.1 repliCATS - Structured elicitations

The repliCATS (Collaborative Assessments for Trustworthy Science) project uses a structured elicitation process--the IDEA protocol--to complete group evaluations of research claims. IDEA stands for: Investigate, Discuss, Estimate, and Aggregate (Figure 3). IDEA is a modified form of the Delphi protocol, with the major differences being that the IDEA protocol encourages interaction between participants and does not require consensus. Interaction between participants takes the form of either face to face discussion or online comments, following evidence that feedback and sharing information improves accuracy of experts' judgments (23), and it sets the IDEA protocol apart from the surveys and prediction markets that have previously been used to predict replicability. In the first half of the program, repliCATS assessments focused on the replicability of research claims. In the remainder of the program, the scope of assessments is expanding to other judgements such as robustness, validity and generalizability. Here we focus on the work from the first half, predicting likely replicability.

Figure 9-3. Overview of the IDEA protocol, as adopted in the repliCATS project



In repliCATS, experts work in small groups of 4 to 6 people using a custom built cloud-based elicitation platform (24,25). Each group is provided with a paper to read and a specific claim from the paper to assess. Individual experts within the group first make their own estimate of whether or not the claim will replicate and document the reasons for their judgement (Investigate). After lodging their initial estimates, individuals receive feedback about their group members’ judgements and reasoning, and they are encouraged to interrogate these and share information (Discuss). Following discussion, each individual provides a second private assessment (Estimate). A mathematical aggregation of the individual estimates is taken as the final assessment (Aggregate). Mathematical aggregation removes the need for group members to reach a consensus.

Mathematical aggregation can take many forms and the repliCATS project has several preregistered aggregation models (<https://osf.io/m6gdp/>). Described in detail by Hanea and colleagues (26), the aggregation models being tested in the repliCATS project fall into three broad categories: (1) linear combinations of best estimates, transformed best estimates (27) and distributions (28); (2) Bayesian approaches, one of

which incorporates characteristics of a claim directly from the paper, such as sample size and effect size; and (3) weighted linear combinations of best estimates, mainly by potential proxies for good forecasting performance, such as demonstrated breadth of reasoning, engagement in the task, openness to changing opinion and informativeness of judgments (29,30). The third category of models is the largest.

The structured elicitation protocol and deliberate inclusion of text responses on the repliCATS platform is fostering an unprecedented qualitative database, with experts documenting the reasoning behind their predictions and judgements. This typically includes justifications for assessments of replicability, and judgements about the papers' importance, clarity and logical structure. The database could increase understanding of how experts evaluate a claim's replicability.

9.1.7.2 Replication markets

The Replication Markets team's approach is motivated by evidence that creative assembly of experts through markets can accurately estimate the replicability of findings in the social and behavioral sciences (1,2,5,6,13,14,31). This approach and evidence build on the well-established ability of markets to aggregate information efficiently (32–36). In a number of contexts (37,38), markets appear to provide better estimates than any one individual can, especially in complex combinatorial prediction markets (39) where individuals make systematic errors (40).

SCORE created two unique challenges for the application of markets: scale and non-resolution. Instead of forecasting replicability of 18-40 similar claims at a time, all

of which would be tested, as has been done in previous replication markets, SCORE required forecasting 3,000 highly diverse claims in about a year, with only a small fraction to be resolved by conducting a replication. We elicited forecasts in 10 monthly rounds of ~300 claims, using a decision market mechanism to preserve proper incentives given the low-resolution rate (Figure 4). Each round of forecasting included replication markets on a set of ~300 claims open for two weeks and a survey for the same set of claims. In replication markets, forecasters traded ‘Yes’ and ‘No’ shares on binary replication questions. ‘Yes’ shares pay 1 point if the replication yields a statistically significant finding in the direction of the original claim. Otherwise ‘No’ shares pay 1 point. The survey directly solicits probabilistic forecasts on replications. A total prize pool was split into one part dedicated to the prediction markets ($\sim\frac{2}{3}$), and one part dedicated to the survey ($\sim\frac{1}{3}$). While the market prizes are paid when replication outcomes become available, the survey prizes were paid each round after the markets closed, using surrogate scores (41) to evaluate each forecaster’s accuracy a month after the round closed when replication outcomes were not yet available. The surrogate scoring method generates a score for a forecast based solely on reported forecasts across claims made by other forecasters. It exploits the unknown statistical correlation of forecasts. Under certain conditions and with enough number of claims and forecasts, it has been theoretically shown that a forecaster’s expected surrogate score reflects their forecast accuracy with respect to the (unavailable) ground truth, and surrogate scoring incentivizes truthful forecasting. For instance, if the Brier score is used to evaluate forecast accuracy against the ground truth, then the surrogate score of a forecaster (without accessing to the ground truth) in expectation equals their Brier score evaluated using the ground truth. Thus, surrogate scoring allows us to provide immediate, potentially noisy, feedback on forecast accuracy before replication outcomes become

available. Once the claim-level replication outcomes are available, we can evaluate forecasting performance in greater detail, similar to the analyses in previous projects. Preregistered tests (42) include the effects of forecaster traits, study features, and aggregation methods on forecast accuracy and outcome. Replication markets and surrogate scoring were also used to forecast the overall replication rate in SCORE and how it depends on research fields and publication year (43).

Figure 9-4. Overview of the Replication Markets workflow.



9.1.8 Machine Assessment

The third technical area (TA3) uses the same dataset of extracted claims to generate confidence scores using machine learning and other algorithmic approaches. The three teams -- PSU, TwoSix, USC -- use different approaches for generating confidence scores.

9.1.8.1 PSU

Researchers at Pennsylvania State University, in collaboration with others at Texas A&M University, Old Dominion University, and Rutgers University use synthetic prediction markets for scoring the replicability of claims. As with the human Replication Market team, a research claim is treated as a binary option in which the

price of the option of a claim at market close can be interpreted as an indicator of confidence in its replicability. Within this framework, artificial agents, or trader-bots, are endowed with initial cash and may choose to purchase options of a given claim, and are trained using an evolutionary algorithm and data from existing replication studies (e.g. (8)).

Prediction markets require the coordinated, sustained effort of collections of human experts limiting their feasibility to scale. Most prediction markets rely on availability of some measurement of ground truth. That is, participants trade on well-defined and verifiable outcomes which are determined after market close. Synthetic prediction markets can overcome these limitations. They can be deployed rapidly and at scale. They can be updated continuously as new information becomes available with periodic, offline human input. Agents can have comprehensive access to prior scholarship far beyond the capacities of an individual researcher. Given the novelty of this approach, the group has dedicated effort to developing a comparable baseline (“Red Team”) led by Texas A&M University and leveraging state of the art approaches for interpretable representation learning developed within DARPA’s XAI program (44–46). Any machine learning (ML) system that can support understanding of the complex factors that contribute to credibility of research claims in practice must explain its outputs. To this end, the complete record of trades, across bots and findings, can offer quantitative understanding of success and failure and provide the basis for learning over time.

In the current functional prototype, asset prices for claims are determined by a logarithmic scoring market rule. Artificial agents are endowed with purchase logic defined using a sigmoid transformation of a convex semi-algebraic set defined in

feature space (47). The team’s feature extraction and representation (FEXRep) framework extracts bibliometric, bibliographic, statistical and semantic features from scientific papers (48–51). So far, 42 distinct features are extracted and provided to bot-traders. To evaluate the bushel claims, the team is expanding feature extraction capabilities, shifting from focusing on paper-level features to incorporate more detailed claim-level features and information about the relationships amongst multiple claims in a single paper. Motivated by a survey of subject matter experts, these features include identifying the theoretical footing of assertions and indicators of rigor in study design.

9.1.8.2 TwoSix

The A+ system developed by Two Six Technologies is a method for understanding replicability given only a journal article in the form of a PDF while encapsulating a wider, more robust set of factors than prior art. The A+ system contains three major computational components: semantic parsing, feature extraction, and replication prediction.

Semantic parsing. The first major step in the A+ system after extracting text from the PDF using Automator is to represent the overall semantic context of each section of text. This is similar to prior annotation work (37,52,53). Here though, we modify the annotation scheme to better match the problem of information extraction for replication prediction (see Table 4). We infer the discourse class for each sentence and perform an averaging of outputs to obtain the final class.

Table 9-4: Discourse classes used in semantic parsing for the A+ method (TwoSix)

Classification	Definition
----------------	------------

Introduction	Problem statement and paper structure
Methodology	Specifics of the study, including participants, materials, and models
Results	Experimental results and statistical tests
Discussion	Author's interpretation of results and implications for the findings
Research Practice	Conflicts of interest, funding sources, and acknowledgements
Reference	Citations

Feature extraction. The unstructured prose of scientific documents includes key features for assessing replicability, such as sample sizes, populations, conditions, experimental variables, methods, materials, exclusion criteria, and participant compensation. Much of this information is available as concise spans of text in the document: “*twenty-four*” may be a sample size; “*undergraduates*” may be a population description; “*reaction time*” may be a dependent variable; and so on. Consequently, we are not interested in extracting and classifying *relations* at this phase of analyses; rather, we optimize our information extractor to classify individual *spans* within the text with context-sensitive labels (e.g., sample count and characteristics, experimental variables, methods), to create a dataset of 620 examples that are annotated with these labels.

Our model next processes the resulting classified spans -- as shown in Figures 5, 6, and 7-- to opportunistically extract domain-specific numerical and Boolean features. For example, the sample count and exclusion count are both expected to be integers, so it attempts to coerce “one hundred and ninety - seven” (Figure 5) and “Eight” (Figure 6) to integers and populate corresponding integer features. Similarly, the model uses a lexicon-based approach over the sample descriptor spans to populate Boolean features indicating whether participants' genders, age, race, religion, and community are specified, what the recruitment pool is (e.g., AMT, universities, etc.), and how they are

compensated (e.g., course credit, monetary, etc.). Because statistical tests are much more structured than each of these features, we use specific Python regular expressions to identify 25 different statistical tests and values including p, R, R², d, F-tests, T-tests, mean, median, standard deviation, confidence intervals, odds ratios, and non-significance.

Figure 9-5: Labeling spans for sample size, sample details, and subject compensation



Figure 9-6: Labeling spans for sample elements excluded and the reason they were excluded

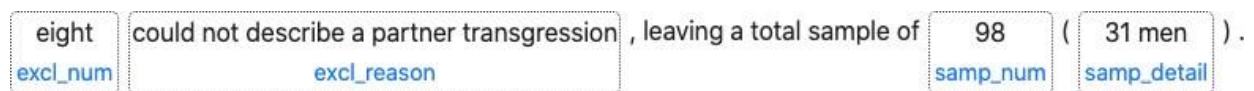
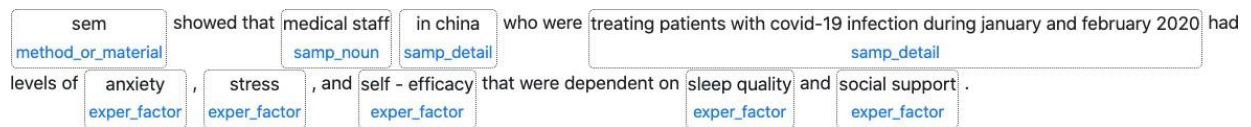


Figure 9-7: Labeling the sample, experimental methods employed, and factors under study



After extracting individual spans and subgraphs from the unstructured prose of a paper, we assemble the extracted information into a global graph called the *argument structure* of the paper. As implied by its name, the argument structure expresses the premises, evidence, and observations in a scientific article, ultimately in support of its conclusions.

The system generates the argument structure by iterating over the sequence of text segments and associated semantic tags to create a structured set of nodes representing the article. For instance, upon encountering a transition in semantic tags,

such as a new **Methodology** section after a **Discussion** section, the system instantiates a new **Study** node and adds the appropriate features.

Replication prediction. The graph-based layout of the argument structure allows the system to assess independent replicability concerns in a context-sensitive, explainable fashion. For example, a sample size of 24 for a study node may impact the judgment of that study's replicability, but it does not necessarily impact the replicability judgment of the study, in the same paper. Each node in the directed argument structure graph is connected directly or indirectly to the node representing the scientific article itself. The argument structure is a fully-connected graph that supports graph and pattern matching, confidence propagation, and feature extraction to judge and explain replicability.

9.1.8.3 University of Southern California

The MACROSCORE system developed by the University of Southern California is a knowledge fusion system that captures a holistic view of the factors important for reproducible and replicable research. The approach mimics the complex judgments that human reviewers make when assessing research. Here, we describe the complex factors and associated techniques for extracting them, the structure and content of the knowledge graph, and the predictive algorithms used in the system. The first pipeline relies on "micro"-features: those that are based on information extracted from papers pertaining to the parameters of the study (e.g., study type and design, sample population as well as indicators of open science, including preregistration, open data, open materials, and open code). Potential detractors to scientific validity, such as conflicts of interest or funding sources are also extracted. To extract these features from

papers, MACROSCORE uses an adaptation of SciBERT, a pre-trained language model created using millions of scientific papers, to identify entities such as experimental parameters, open science indicators, and claim information. Together, these provide a core set of document-specific features.

The second pipeline in MACROSCORE is the "macro"-feature pipeline that captures the broader scientific context of a paper. Determining the impact and contributions of a scientific work is a difficult and subjective task. MACROSCORE addresses these challenges by applying network science approaches to the bibliometric structure of scientific disciplines. Specifically, MACROSCORE collects the citations and references within a particular scientific discipline, forming a network connecting the scientific articles and their authors. Metrics of network structure, including in-degree (incoming citations to the work), out-degree (references to other works), authority score (citations by important works), and hub score (citing important work) provide core features to assess the scientific work.

The heart of the MACROSCORE system is a knowledge graph that represents the features distilled from both micro and macro pipelines. The knowledge graph represents the core concepts of the scientific discipline: scholarly works, scientific claims, scholars, organizations, and publication venues. MACROSCORE uses an ontology derived from the popular, public, and widely-used knowledge graph Wikidata to include each scientific article, the journal where it was published, its authors and editors, and the affiliations of each, and all citations and references to the article. Beyond the classes and properties defined in Wikidata, MACROSCORE has extended the ontology on Wikidata to incorporate claim information described earlier, as well as derived features from four high-level classes: validity of inference, study design,

reporting and transparency, and scientific network. Together, these features create a comprehensive profile of the scientific work and its connection to other works.

The final component of the MACROSCORE system is a suite of predictive algorithms that operate on the features from each pipeline and the knowledge graph. Among other methods, MACROSCORE uses a probabilistic graphical model using the probabilistic soft logic (PSL) framework. This model includes dependencies between different features defined in the knowledge graph specified as logical rules, such as "Small sample sizes and small effect sizes indicate poor replicability." Using training data, the PSL framework can learn the importance of each rule and its associated features. For a given judgment made by the MACROSCORE system, the PSL model will provide a set of explanatory statements, and an analysis of the top features contributing to the assessment. As the system evolves, MACROSCORE will incorporate more features from both the article and scientific network, and create an increasingly comprehensive knowledge graph.

9.1.9 Empirical evidence for credibility assessment

Independent empirical assessments provide the basis for evaluating the confidence scores generated by humans and algorithms to predict credibility of claims. Table 5 presents approaches to empirical assessment of credibility roughly ordered from the bottom being the least effortful but providing the least information about credibility to the top being the most effortful and providing the most information about credibility. "Roughly" is an important qualifier because there are many exceptions based on particular cases for which amount of effort and amount of information may not

correspond cleanly with this depiction. In general, lower categories in Table 5 correspond with assessments of the original design and original data for a narrow test of whether the original report found what it reported to have found, and higher categories correspond with more laborious assessments involving obtaining new designs and data for a broader test of whether the original claim is supported by new evidence. These are not the only ways to assess credibility. For example, a finding could be reproducible, robust, replicable, generalizable, and invalid. Nevertheless, these assessments are tractable and verifiable indicators that are related to other aspects of credibility.

As TA1, COS bears responsibility for coordinating a large network of social-behavioral researchers to contribute empirical evidence assessing the credibility of claims. The team draws on the stratified random sample of 600 claims comprising the *evidence set* and matches their topics and methodologies to researchers with appropriate resources and expertise to conduct an empirical assessment. The focus of the first half of the SCORE program was on conducting replication and reproduction studies. The remainder of the program expands the scope of empirical evidence to include all of the forms presented in Table 5.

Table 9-5. Forms of empirical credibility assessment

Generalizable	Original claim supported across diverse samples, treatments, outcomes, and settings
Replicable	Original claim supported with independent evidence
Robust	Original claim supported with diverse treatments of original data
Outcome Reproducible	Original claim supported with original analysis of original data
Process Reproducible	Possible to assess outcome reproducibility of original claim

Internally consistent	Reporting of original claim does not have detectable errors
-----------------------	-------------------------------------------------------------

A reproduction refers to applying the original analysis strategy to the original data to test whether the same result recurs. A reproduction could fail due to process reproducibility because, for example, the original data are not available, making it impossible to conduct the analysis again. This does not disconfirm the original finding, but it is a credibility risk in that the original finding cannot be confirmed or disconfirmed. A reproduction could also fail due to outcome reproducibility because, for example, applying the analysis described in the original paper does not produce the finding associated with it. This can occur because of errors in reporting, ambiguity in description of analyses, or factors in the data analysis pipeline.

A replication refers to testing the original claim with different data. That data could be pre-existing, such as re-testing the relationship between variables in a subsequent wave of a panel study, or could be newly generated with a study design to test the same research question. Whether based on existing or new data, the determination of whether a new test is a replication of a prior claim is a theoretical commitment that the inevitable differences between the original and replication study are irrelevant for testing the original claim (54).

To provide evidence that is both appropriate to testing individual claims and standard enough to evaluate SCORE teams' prediction methods across disciplines, we designed a process that balances specific requirements that all projects must adhere to with ongoing evaluation and feedback by subject area experts. For example, all replications are prepared using a standard template that is reviewed by 2-3 independent

researchers, and the resolution of design changes suggested by reviewers is managed by an editor. Authors of the original finding are invited to participate in the review process or to submit a commentary on the design. The review process is intended to improve the quality of the replication designs so that they are effective, good-faith tests of the original claim. The template and review process also provide an occasion to explicitly document differences between original and replication studies and assessments of any heterogeneity in beliefs about whether they are consequential for the replication design.

Following approval, the design and analysis plan is preregistered on the Open Science Framework (OSF). Research teams conduct their studies and then report outcomes following a standard protocol and provide all research materials, data, and code so that the replication studies are themselves reproducible and, eventually, accessible to others to the extent ethically possible. The reproduction workflow has a similar emphasis on documentation and transparency with a lighter review process emphasizing adherence to the standardized protocol for reproducing original findings.

As singular attempts to reproduce or replicate original claims, these empirical efforts do not provide definitive evidence about their credibility (8) -- they add to the body of evidence about that claim which includes the original paper and any other evidence for the claim in the literature. However, prior evidence that both humans and algorithms can predict the outcomes of these reproductions and replications provides a basis for treating them as ground truth for the purposes of the program. More importantly, the generated dataset of original and novel statistical evidence, reproduction and replication outcomes, along with the expanded set of empirical credibility indicators from internal consistency (e.g., statcheck), robustness (e.g., multiverse or many-analyst investigations), or generalizability tests will provide a rich

network of evidence to investigate convergence and heterogeneity of these credibility indicators.

9.1.10 Evaluating Expert and Machine Success

There is no definitive criterion for deciding whether a finding is successfully replicated or reproduced (9), but pragmatic, defensible, and widely applicable benchmarks are needed to evaluate the outcomes of the SCORE program. The role of the MITRE Testing & Evaluation (T&E) team in SCORE is to evaluate the relative match between predicted and actual confidence in each claim using the outcomes from the TA1 empirical results and the human-generated confidence scores from TA2. T&E focuses on evaluating the accuracy of human-generated confidence scores relative to replication outcomes and the accuracy of algorithm-generated confidence scores relative to the most accurate human-generated scores.

Evaluation of human confidence score accuracy against binary replication outcomes focuses on discrimination or “signal detection”(55) – that is, the ability to prospectively distinguish claims with higher and lower chances of successful replication on the basis of reliably diagnostic indicators. In addition to a modified version of the Wilcoxon Mann-Whitney U statistic (56), we use an area under the curve (AUC) interpretation which can be understood as the “meta-probability” that the forecast system assigns a higher probability to a “positive” case than to a “negative” case for any randomly sampled pairing of two such cases (57,58).

The analysis of replication p-values are used as a supplementary continuous measure of a claim’s degree or amount of replication success, where smaller replication p-values indicate higher levels of replication study support for the original study claim.

Additional supplementary metrics include: stand-alone reporting of proper scoring rule values (59), measures of calibration (60), and various “confusion matrix”-style measures of classification performance (e.g., sensitivity, specificity, proportionate reduction in error vs. base rate; (57)). Using metrics based on the p-value to assess replication outcomes have known limitations (8). However, they also have the virtues of easy application, straightforward interpretability, broad applicability across research methodologies, and demonstrated validity in prior human and machine prediction contexts (1,2,13,14,16,18).

To evaluate algorithm accuracy in predicting human confidence scores, the root mean squared error (RMSE) is used as one of two primary outcome metrics. Additionally, Kendall’s tau-b, a nonparametric measure of monotonic association (56) is used to assess accuracy in discriminating among claims with greater or lesser amounts of replication support. Finally, we use measures of calibration as a supplementary metric (e.g., regression of TA2 scores on TA3 scores, where intercept and slope deviating from 0 and 1, respectively, would be evidence of miscalibration).

Finally, toward the end of the SCORE program, RAND researchers will pilot the use of TA3 tools to assess their applicability with users in the policy community. While few studies have an explicit emphasis on the reproducibility of scientific claims, matters of generalization and reliability weigh heavily on the development and assessment of policy interventions. Two applications of particular interest include the ability to characterize findings from large bodies of literature that form the initial basis of information from which further studies are drawn, and in the role of adjudicating load-bearing claims that may be sources of contention among policy making stakeholders.

9.1.11 Potential Outcomes, Findings, and Artifacts

The primary research objective for SCORE is to create accurate, scalable, automated algorithms to signal confidence in research claims. There are a variety of potential use cases. Researchers might use scores to identify potential weaknesses in their claims and provide more detail or support. Journal editors and conference organizers might use the scores to prioritize selection of reviewers with expertise in areas that the algorithm flagged as low confidence.

Funders and researchers designing proposals might use the scores to identify potentially important findings that have not yet achieved high confidence. The scores could guide policymakers' information search and allocation of effort to obtain additional evidence or expert judgment when the algorithm flags uncertainty.

Across use cases, such a technology would provide a heuristic “first pass” to help direct attention to areas of risk and opportunity. To be clear, even the most optimistic assessments of the potential of such scores would not defer reasoning, decision-making, judgment, and action to machines. As in other applications, uncritical use of algorithms can perpetuate biases in how we evaluate claims, or reflect inappropriate generalizations about what signals indicate that a paper is credible (61–63).

Effective automated technologies can be a tool to complement these human and social processes in the assessment, prioritization, and application of research. They can also provide researchers with tools for rapid and iterative assessments of credibility. At scale, as an iterative feedback mechanism, they may help foster culture and behavioral changes that increase the overall credibility of research.

SCORE represents a unique opportunity to explore a challenge that is paramount to modern AI--How can we combine the best of both human and machine reasoning? The nuance inherent in scientific expression beyond the obvious reporting of statistical information makes this program both challenging and exciting. Explainability of results in machine learning is always challenging, but made more so by the complex environment of human writing. With multiple algorithm strategies using enriched extracted information from papers and human judgment and replication outcomes as validation measures, SCORE may facilitate significant progress on this problem.

Beyond the primary objectives, SCORE will advance a variety of research questions about the credibility and assessment of scholarly research, and generate research artifacts that can support dozens or hundreds of investigations. These artifacts include:

- *Annotation Set*: A stratified random sample of 3,000 papers with a claim trace from the abstract to a statistical inference in the paper from a stratified random sample of about 30,000 papers from >60 journals from the social-behavioral sciences from 2009 to 2018 with metadata enhancements such as open science badges, links to open access versions of articles, and code availability statements;
- *Confidence scores*: Expert and machine ratings of the confidence in *Annotation Set* claims along with substantial metadata and qualitative assessments about the papers and basis for confidence ratings;
- *Evidence set*: A stratified random sample of 600 papers from the *Annotation set* that additionally assess statistical errors in the papers, process and outcome reproducibility, robustness, and/or replicability;

- *Enhanced bushel set*: After 200 of the 600 papers undergo further enhancement by extracting a full bushel of claims tracing from the abstract to statistical inferences in the paper, experts and machines will provide scores and other assessments of all claims, and some additional reproduction, robustness, and replication evidence will be accumulated for multiple claims in those papers;
- *Process data and artifacts from project execution*: Substantial data and documentation about the process of conducting this work and the many additional artifacts that are created along the way, sufficient to extend the artifacts and make it a living body of research. Cumulatively, SCORE is the most in-depth examination of credibility of research claims in the social and behavioral sciences ever conducted.

All of the data and materials from SCORE that can be shared without violating publisher intellectual property rights or human participant protections will be made publicly accessible after the program is completed. There are many possible research questions that will be possible to advance with these data by any interested researchers. For example, some of the questions that the SCORE team is already investigating with these data include: What is the strength of evidence in original claims? How do experts and machines evaluate the credibility of claims and how does this vary by discipline, time, topic, and methodology? What are observed reproducibility, robustness, and replicability rates in the sample and how do they likewise vary? How well do humans and machines predict replicability, robustness, and reproducibility? How are credibility indicators related to one another?

9.1.12 Conclusion

SCORE has aspirational objectives to advance scalable tools for credibility assessment, and will generate substantial research artifacts to support scholarly research on human and machine judgment, replicability and reproducibility, and the nature of research claims. This is made possible by SCORE's greatest asset -- the participation of hundreds of researchers across the social and behavioral sciences that are contributing to claim extraction, credibility assessment, and reproducibility, robustness, and replication studies. This large-scale team science project is generating data that would not otherwise be possible (64), and will open doors to many novel investigations to assess and enhance research credibility. If nothing else, the program may provide a case example of the potential for team science in tackling many of the most important challenges in social and behavioral research.

9.1.13 References

1. Camerer CF, Dreber A, Forsell E, Ho T-H, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. *Science*. 2016 Mar 25;351(6280):1433–6.
2. Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in *Nature and Science* between 2010 and 2015. *Nat Hum Behav*. 2018 Sep;2(9):637–44.
3. Cova F, Strickland B, Abatista A, Allard A, Andow J, Attie M, et al. Estimating the Reproducibility of Experimental Philosophy. *Rev Philos Psychol*. 2021 Mar 1;12(1):9–44.
4. Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, Banks JB, et al. Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J Exp Soc Psychol*. 2016 Nov 1;67:68–82.

5. Ebersole CR, Mathur MB, Baranski E, Bart-Plange D-J, Buttrick NR, Chartier CR, et al. Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. *Adv Methods Pract Psychol Sci*. 2020 Sep 1;3(3):309–31.
6. Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB, Alper S, et al. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Adv Methods Pract Psychol Sci*. 2018 Dec;1(4):443–90.
7. Klein RA, Ratliff KA, Vianello M, Adams Jr. RB, Bahník Š, Bernstein MJ, et al. Investigating variation in replicability: A “many labs” replication project. *Soc Psychol*. 2014;45(3):142–52.
8. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015 Aug 28;349(6251):aac4716.
9. Nosek BA, Hardwicke TE, Moshontz H, Allard A, Corker KS, Almenberg AD, et al. Replicability, Robustness, and Reproducibility in Psychological Science *PsyArXiv*; 2021 [cited 2021 Mar 25]. Available from: <https://psyarxiv.com/ksfvq/>
10. Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*. 2020 Jun;582(7810):84–8.
11. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Adv Methods Pract Psychol Sci*. 2018 Sep 1;1(3):337–56.
12. Simonsohn U, Simmons JP, Nelson LD. Specification curve analysis. *Nat Hum Behav*. 2020 Nov;4(11):1208–14.
13. Forsell E, Viganola D, Pfeiffer T, Almenberg J, Wilson B, Chen Y, et al. Predicting replication outcomes in the Many Labs 2 study. *J Econ Psychol*. 2019 Dec 1;75:102117.
14. Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, et al. Using prediction markets to estimate the reproducibility of scientific research. *Proc Natl Acad Sci*. 2015 Dec 15;112(50):15343–7.
15. Wintle B, Mody F, Smith E, Hanea A, Wilkinson DP, Hemming V, et al. Predicting and reasoning about replicability using structured groups *MetaArXiv*; 2021 [cited 2021 Sep 23]. Available from: <https://osf.io/preprints/metaarxiv/vtpmb/>
16. Altmejd A, Dreber A, Forsell E, Huber J, Imai T, Johannesson M, et al. Predicting the replicability of social science lab experiments. *PLOS ONE*. 2019 May 12;14(12):e0225826.
17. Pawel S, Held L. Probabilistic forecasting of replication studies. *PLOS ONE*. 2020 Apr 22;15(4):e0231416.

18. Yang Y, Youyou W, Uzzi B. Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proc Natl Acad Sci*. 2020 May 19;117(20):10762–8.
19. Chang AC, Li P. *Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say “Usually Not”* Rochester, NY: Social Science Research Network; 2015 Sep [cited 2021 Sep 23]. Report No.: ID 2669564. Available from: <https://papers.ssrn.com/abstract=2669564>
20. Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA. An open investigation of the reproducibility of cancer biology research. Rodgers P, editor. *eLife*. 2014 Dec 10;3:e04333.
21. McCullough B d., McGeary KA, Harrison TD. Do economics journal archives promote replicable research? *Can J Econ Can Déconomique*. 2008;41(4):1406–20.
22. Wood BDK, Müller R, Brown AN. Push button replication: Is impact evaluation evidence for international development verifiable? *PLOS ONE*. 2018 Dec 21;13(12):e0209416.
23. Kerr NL, Tindale RS. Group Performance and Decision Making. *Annu Rev Psychol*. 2004;55(1):623–55.
24. Fraser H, Bush M, Wintle B, Mody F, Smith E, Hanea A, et al. Predicting reliability through structured expert elicitation with repliCATS (Collaborative Assessments for Trustworthy Science) *MetaArXiv*; 2021 [cited 2021 Sep 23]. Available from: <https://osf.io/preprints/metaarxiv/2pczv/>
25. Pearson R, Fraser H, Bush M, Mody F, Widjaja I, Head A, et al. Eliciting group judgements about replicability: a technical implementation of the IDEA Protocol 2021 [cited 2021 Sep 23]. Available from: <http://scholarspace.manoa.hawaii.edu/handle/10125/70666>
26. Hanea A, Wilkinson DP, McBride M, Lyon A, Ravenzwaaij D van, Thorn FS, et al. Mathematically aggregating experts’ predictions of possible futures *MetaArXiv*; 2021 [cited 2021 Sep 23]. Available from: <https://osf.io/preprints/metaarxiv/rxmh7/>
27. Satopää VA, Baron J, Foster DP, Mellers BA, Tetlock PE, Ungar LH. Combining multiple probability predictions using a simple logit model. *Int J Forecast*. 2014 Apr 1;30(2):344–56.
28. Cooke RM, Marti D, Mazzuchi T. Expert forecasting with and without uncertainty quantification and weighting: What do the data say? *Int J Forecast*. 2021 Jan 1;37(1):378–87.
29. Mellers B, Stone E, Atanasov P, Rohrbaugh N, Metz SE, Ungar L, et al. The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *J Exp Psychol Appl*. 2015;21(1):1–14.

30. Mellers B, Stone E, Murray T, Minster A, Rohrbaugh N, Bishop M, et al. Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspect Psychol Sci*. 2015 May 1;10(3):267–81.
31. Gordon M, Viganola D, Dreber A, Johannesson M, Pfeiffer T. Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. *PLOS ONE*. 2021 Apr 14;16(4):e0248780.
32. Arrow KJ, Forsythe R, Gorham M, Hahn R, Hanson R, Ledyard JO, et al. The Promise of Prediction Markets. *Science*. 2008 May 16;320(5878):877–8.
33. Malkiel BG, Fama EF. Efficient Capital Markets: A Review of Theory and Empirical Work*. *J Finance*. 1970;25(2):383–417.
34. Plott CR, Wit J, Yang WC. Parimutuel Betting Markets as Information Aggregation Devices: Experimental Results. *Econ Theory*. 2003;22(2):311–51.
35. Plott CR, Sunder S. Rational Expectations and the Aggregation of Diverse Information in Laboratory Security Markets. *Econometrica*. 1988;56(5):1085–118.
36. Radner R. Rational Expectations Equilibrium: Generic Existence and the Information Revealed by Prices. *Econometrica*. 1979;47(3):655–78.
37. Chen K-Y, Fine LR, Huberman BA. Predicting the Future. *Inf Syst Front*. 2003 Jan 1;5(1):47–61.
38. Forsythe R, Rietz TA, Ross TW. Wishes, expectations and actions: a survey on price formation in election stock markets. *J Econ Behav Organ*. 1999 May 1;39(1):83–110.
39. Chen Y, Pennock DM. Designing Markets for Prediction. *AI Mag*. 2010 Dec;31(4):42–52.
40. Wang G, Kulkarni SR, Poor HV, Osherson DN. Aggregating Large Sets of Probabilistic Forecasts by Weighted Coherent Adjustment. *Decis Anal*. 2011 Jun 1;8(2):128–44.
41. Liu Y, Wang J, Chen Y. Surrogate Scoring Rules. *ACM Conf Econ Comput* 2020 Jul; Available from: <http://arxiv.org/abs/1802.09158>
42. Pfeiffer T, Chen Y, Viganola D, Bishop M, Almenberg AD, Johannesson M, et al. Prereg Replication Markets Rounds 1-10 (Version 7). 2020 Oct 6 [cited 2021 Sep 23]; Available from: <https://osf.io/svg3x>
43. Gordon M, Viganola D, Bishop M, Chen Y, Dreber A, Goldfedder B, et al. Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *R Soc Open Sci*. 2020 Jul 22;7(7):200566.
44. Du M, Liu N, Song Q, Hu X. Towards Explanation of DNN-based Prediction with Guided Feature Inversion. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* New York,

- NY, USA: Association for Computing Machinery; 2018 [cited 2021 Sep 23]. p. 1358–67. (KDD '18). Available from: <https://doi.org/10.1145/3219819.3220099>
45. Du M, Liu N, Yang F, Hu X. Learning credible DNNs via incorporating prior knowledge and model local explanation. *Knowl Inf Syst*. 2021 Feb 1;63(2):305–32.
 46. Yang F, Liu N, Wang S, Hu X. Towards Interpretation of Recommender Systems with Sorted Explanation Paths. In: 2018 IEEE International Conference on Data Mining (ICDM). 2018. p. 667–76.
 47. Nakshatri N, Menon A, Giles CL, Rajtmajer S, Griffin C. Design and analysis of a synthetic prediction market using dynamic convex sets. *Results Control Optim*. 2021 Sep 11;100052.
 48. Lanka SST, Rajtmajer S, Wu J, Giles CL. Extraction and Evaluation of Statistical Information from Social and Behavioral Science Papers. In: Companion Proceedings of the Web Conference 2021 New York, NY, USA: Association for Computing Machinery; 2021 [cited 2021 Sep 23]. p. 426–30. (WWW '21). Available from: <https://doi.org/10.1145/3442442.3451363>
 49. Modukuri SA, Rajtmajer S, Squicciarini AC, Wu J, Giles CL. Understanding and predicting retractions of published work: 2021 Workshop on Scientific Document Understanding, SDU 2021. *CEUR Workshop Proc* 2021 [cited 2021 Sep 24];2831. Available from: <http://www.scopus.com/inward/record.url?scp=85103079212&partnerID=8YFLogxK>
 50. Wu J, Nivargi R, Lanka SST, Menon AM, Modukuri SA, Nakshatri N, et al. Predicting the Reproducibility of Social and Behavioral Science Papers Using Supervised Learning Models. *ArXiv210404580 Cs* 2021 Apr 7 [cited 2021 Sep 23]; Available from: <http://arxiv.org/abs/2104.04580>
 51. Wu J, Wang P, Wei X, Rajtmajer S, Giles CL, Griffin C. Acknowledgement Entity Recognition in COVID-19 Papers. In: Proceedings of the First Workshop on Scholarly Document Processing Online: Association for Computational Linguistics; 2020 [cited 2021 Sep 23]. p. 10–9. Available from: <https://aclanthology.org/2020.sdp-1.3>
 52. Dasigi P, Burns GAPC, Hovy E, de Waard A. Experiment Segmentation in Scientific Discourse as Clause-level Structured Prediction using Recurrent Neural Networks. *ArXiv170205398 Cs* 2017 Feb 17 [cited 2021 Sep 23]; Available from: <http://arxiv.org/abs/1702.05398>
 53. Huber P, Carenini G. Predicting Discourse Structure using Distant Supervision from Sentiment. *ArXiv191014176 Cs* 2019 Oct 30 [cited 2021 Sep 23]; Available from: <http://arxiv.org/abs/1910.14176>
 54. Nosek BA, Errington TM. What is replication? *MetaArXiv*; 2019 Sep [cited 2019 Oct 15]. Available from: <https://osf.io/u4g6t>

55. Yaniv I, Yates JF, Smith JK. Measures of discrimination skill in probabilistic judgment. *Psychol Bull.* 1991;110(3):611.
56. Gibbons JD. *Nonparametric measures of association.* Thousand Oaks, CA, US: Sage Publications, Inc; 1993. vi, 97 p. (Nonparametric measures of association).
57. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press; 2003. 319 p.
58. Steyvers M, Wallsten TS, Merkle EC, Turner BM. Evaluating Probabilistic Forecasts with Bayesian Signal Detection Models. *Risk Anal.* 2014;34(3):435–52.
59. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev.* 1950 Jan 1;78(1):1–3.
60. Arkes HR, Dawson NV, Speroff T, Harrell FE, Alzola C, Phillips R, et al. The covariance decomposition of the probability score and its use in evaluating prognostic estimates. *SUPPORT Investigators. Med Decis Making.* 1995 Apr 1;15(2):120–31.
61. Buolamwini J, Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Friedler SA, Wilson C, editors. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency PMLR;* 2018. p. 77–91. (Proceedings of Machine Learning Research; vol. 81). Available from: <https://proceedings.mlr.press/v81/buolamwini18a.html>
62. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science.* 2017 Apr 14;356(6334):183–6.
63. Larson J, Mattu S, Kirchner L, Angwin J. How we analyzed the COMPAS recidivism algorithm. *ProPublica* 5 2016. 2016;9(1).
64. Uhlmann EL, Ebersole CR, Chartier CR, Errington TM, Kidwell MC, Lai CK, et al. Scientific Utopia III: Crowdsourcing Science. *Perspect Psychol Sci.* 2019 Sep 1;14(5):711–33.

9.2 Appendix 2: DRC16 Forms

DRC 16



GRADUATE
RESEARCH
SCHOOL

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Michael Gordon
Name/title of Primary Supervisor:	Professor Thomas Pfeiffer
In which chapter is the manuscript /published work:	Chapter 2
Please select one of the following three options:	
<input checked="" type="radio"/> The manuscript/published work is published or in press <ul style="list-style-type: none"> • Please provide the full reference of the Research Output: Gordon M, Viganola D, Dreber A, Johannesson M, Pfeiffer T (2021) Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. PLoS ONE 16(4): e0248780. https://doi.org/10.1371/journal.pone.0248780 	
<input type="radio"/> The manuscript is currently under review for publication – please indicate: <ul style="list-style-type: none"> • The name of the journal: • The percentage of the manuscript/published work that was contributed by the candidate: • Describe the contribution that the candidate has made to the manuscript/published work: 	
<input type="radio"/> It is intended that the manuscript will be published, but it has not yet been submitted to a journal	
Candidate's Signature:	
Date:	30/10/2020
Primary Supervisor's Signature:	Pfeiffer, Thomas <small>Digitally signed by Pfeiffer, Thomas Date: 2021.09.30 16:35:27 +1300'</small>
Date:	30-Oct-2020

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.


GRS Version 5 – 13 December 2019
DRC 19/09/10



GRADUATE
RESEARCH
SCHOOL

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Michael Gordon
Name/title of Primary Supervisor:	Professor Thomas Pfeiffer
In which chapter is the manuscript /published work:	Chapter 4
Please select one of the following three options:	
<input checked="" type="radio"/> The manuscript/published work is published or in press <ul style="list-style-type: none"> • Please provide the full reference of the Research Output: Gordon, M., Viganola, D., Bishop, M., Chen, Y., Dreber, A., Goldfedder, B., ... & Pfeiffer, T. (2020). Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. Royal Society open science. 	
<input type="radio"/> The manuscript is currently under review for publication – please indicate: <ul style="list-style-type: none"> • The name of the journal: • The percentage of the manuscript/published work that was contributed by the candidate: • Describe the contribution that the candidate has made to the manuscript/published work: 	
<input type="radio"/> It is intended that the manuscript will be published, but it has not yet been submitted to a journal	
Candidate's Signature:	
Date:	30/10/2020
Primary Supervisor's Signature:	Pfeiffer, Thomas <small>Digitally signed by Pfeiffer, Thomas Date: 2021.09.30 16:36:07 +13'00'</small>
Date:	30-Oct-2020


This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.



GRADUATE
RESEARCH
SCHOOL

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS


We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Michael Gordon
Name/title of Primary Supervisor:	Professor Thomas Pfeiffer
In which chapter is the manuscript /published work: Chapter 5	
Please select one of the following three options:	
<input checked="" type="radio"/> The manuscript/published work is published or in press <ul style="list-style-type: none"> • Please provide the full reference of the Research Output: Tierney, Warren, Jay H. Hardy III, Charles R. Ebersole, Keith Leavitt, Domenico Viganola, Elena Giulia Clemente, Michael Gordon et al. "Creative destruction in science." <i>Organizational Behavior and Human Decision Processes</i> 161 (2020): 291-309. 	
<input type="radio"/> The manuscript is currently under review for publication – please indicate: <ul style="list-style-type: none"> • The name of the journal: • The percentage of the manuscript/published work that was contributed by the candidate: • Describe the contribution that the candidate has made to the manuscript/published work: 	
<input type="radio"/> It is intended that the manuscript will be published, but it has not yet been submitted to a journal	
Candidate's Signature:	
Date:	30/10/2020
Primary Supervisor's Signature:	Pfeiffer, Thomas <small>Digitally signed by Pfeiffer, Thomas Date: 2021.09.30 16:36:43 +13'00'</small>
Date:	30-Oct-2020

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Michael Gordon
Name/title of Primary Supervisor:	Professor Thomas Pfeiffer
In which chapter is the manuscript /published work:	Chapter 6
<p>Please select one of the following three options:</p> <p><input checked="" type="radio"/> The manuscript/published work is published or in press</p> <ul style="list-style-type: none"> • Please provide the full reference of the Research Output: <p>Tierney, W., Hardy III, J., Ebersole, C. R., Viganola, D., Clemente, E. G., Gordon, M., ... & Culture & Work Morality Forecasting Collaboration. (2021). A creative destruction approach to replication: Implicit work and sex morality across cultures. <i>Journal of Experimental Social Psychology</i>, 93, 104060.</p> <p><input type="radio"/> The manuscript is currently under review for publication – please indicate:</p> <ul style="list-style-type: none"> • The name of the journal: • The percentage of the manuscript/published work that was contributed by the candidate: • Describe the contribution that the candidate has made to the manuscript/published work: <p><input type="radio"/> It is intended that the manuscript will be published, but it has not yet been submitted to a journal</p>	
Candidate's Signature:	
Date:	30/10/2020
Primary Supervisor's Signature:	Pfeiffer, Thomas <small>Digitally signed by Pfeiffer, Thomas Date: 2021.09.30 16:37:06 +13'00'</small>
Date:	30-Oct-2020


This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.



GRADUATE
RESEARCH
SCHOOL

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Michael Gordon
Name/title of Primary Supervisor:	Professor Thomas Pfeiffer
In which chapter is the manuscript /published work: Chapter 7	
Please select one of the following three options:	
<input type="radio"/> The manuscript/published work is published or in press <ul style="list-style-type: none"> • Please provide the full reference of the Research Output: 	
<input type="radio"/> The manuscript is currently under review for publication – please indicate: <ul style="list-style-type: none"> • The name of the journal: • The percentage of the manuscript/published work that was contributed by the candidate: • Describe the contribution that the candidate has made to the manuscript/published work: 	
<input checked="" type="radio"/> It is intended that the manuscript will be published, but it has not yet been submitted to a journal	
Candidate's Signature:	
Date:	30/10/2020
Primary Supervisor's Signature:	Pfeiffer, Thomas <small>Digitally signed by Pfeiffer, Thomas Date: 2021.09.30 16:37:33 +13'00'</small>
Date:	30-Oct-2020

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.