Systematizing Confidence in Open Research and Evidence (SCORE)

SCORE Collaboration[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

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.

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 (Camerer et al., 2016, 2018; Cova et al., 2018; Ebersole et al., 2016, 2020; Klein et al., 2014, 2018; Open Science Collaboration, 2015). For example, Nosek and colleagues (2021) 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 (Botvinik-Nezer et al., 2020; Silberzahn et al., 2018; Simonsohn et al., 2020). 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 (Camerer et al., 2016, 2018; Dreber et al., 2015; Forsell et al., 2019; Wintle et al., 2021). 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., 2021 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 (Altmejd et al., 2019; Pawel & Held, 2020; Y. Yang et al., 2020). 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.

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 1. Journals comprising the Common Task Framework (CTF)

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

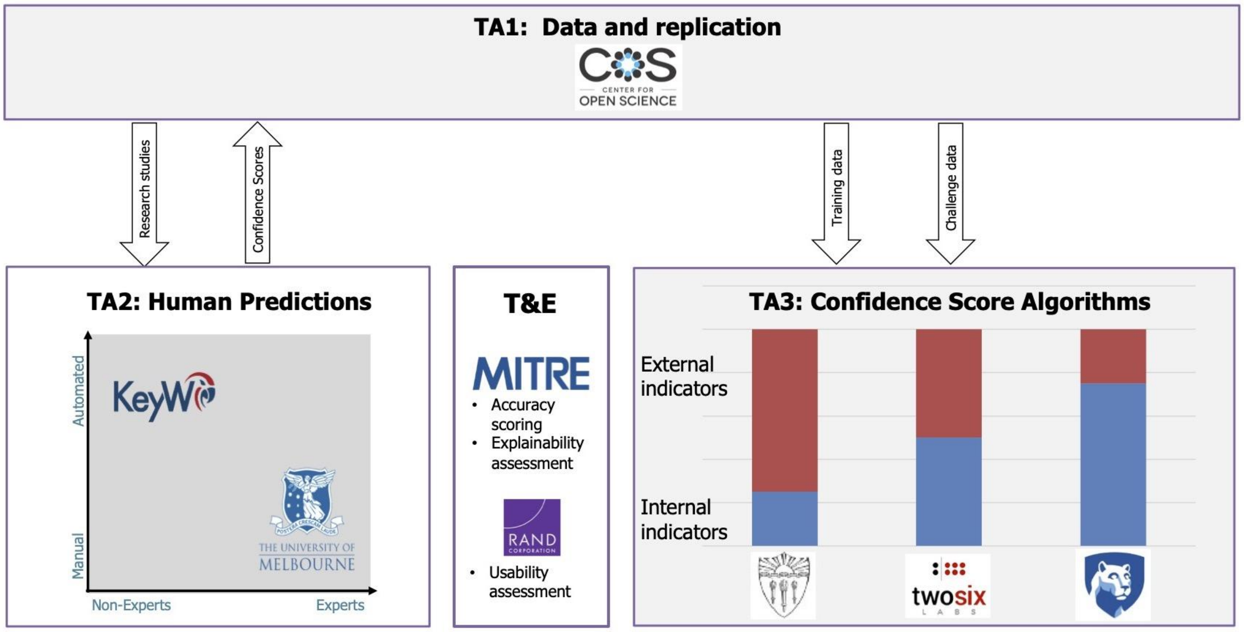
	Journal of Conflict Resolution *Comparative Political Studies* *World Development*
Psychology	*Journal of Experimental Social Psychology* *Journal of Applied Psychology* *Journal of Environmental Psychology* *Journal of Personality and Social Psychology* *Journal of Experimental Psychology: General*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

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 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.

What Makes SCORE Unique

SCORE draws inspiration from prior research on systematic replications and reproductions (Camerer et al., 2016, 2018; Chang & Li, 2015; Cova et al., 2018; Ebersole et al., 2016, 2020; Errington et al., 2014; Klein et al., 2014, 2018; McCullough et al., 2008; Open Science Collaboration, 2015; Wood et al., 2018) and replicability predictions by humans (Camerer et al., 2018; Dreber et al., 2015; Forsell et al., 2019) and machines (Altmejd et al., 2019; Pawel & Held, 2020; Y. Yang et al., 2020). 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.

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 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.

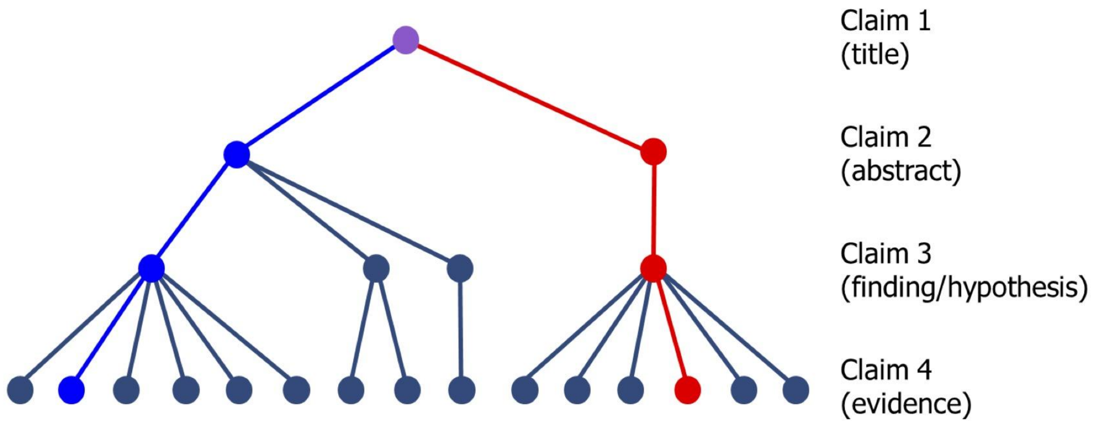
Evidence Set	A stratified random sample of approximately 600 papers from the annotation set. These papers could be selected for an empirical attempt to find further evidence for or against a claim.
Claim	A specific assertion reported as a finding in a paper. Most papers make more than one claim, and 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 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 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.

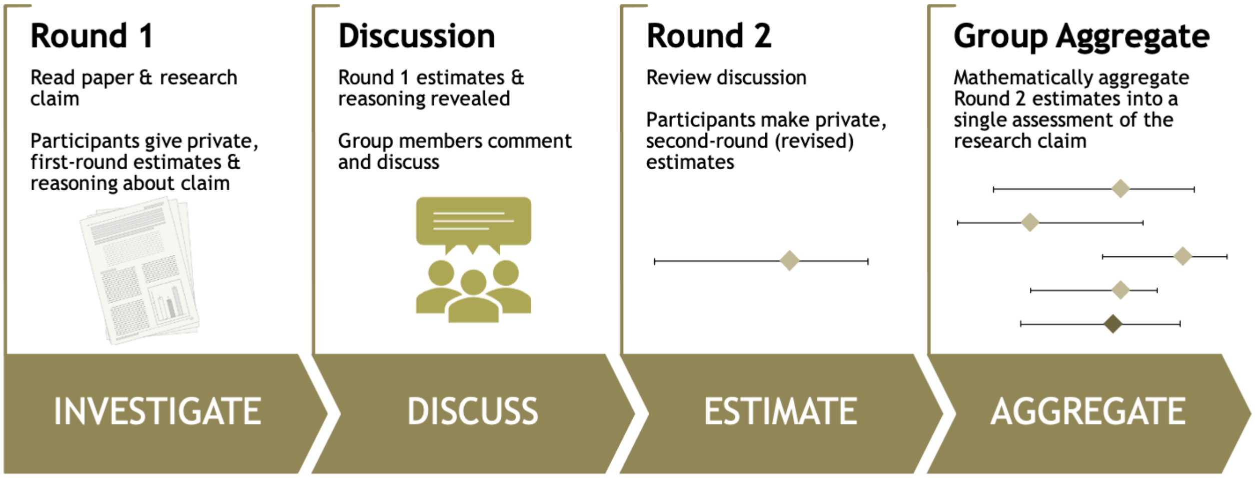
Expert Assessment

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.

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 (Kerr & Tindale, 2004), 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 3. Overview of the IDEA protocol, as adopted in the repliCATS project



Round 1

Read paper & research claim

Participants give private, first-round estimates & reasoning about claim

Discussion

Round 1 estimates & reasoning revealed

Group members comment and discuss

Round 2

Review discussion

Participants make private, second-round (revised) estimates

Group Aggregate

Mathematically aggregate Round 2 estimates into a single assessment of the research claim

In repliCATS, experts work in small groups of 4 to 6 people using a custom built cloud-based elicitation platform (Fraser et al., 2021; Pearson et al., 2021). 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 (2021), the aggregation models being tested in the repliCATS project fall into three broad categories: (1) linear combinations of best estimates, transformed best estimates (Satopää et al., 2014) and distributions (Cooke et al., 2021); (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 (Mellers, Stone, Atanasov, et al., 2015; Mellers, Stone, Murray, et al., 2015). 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.

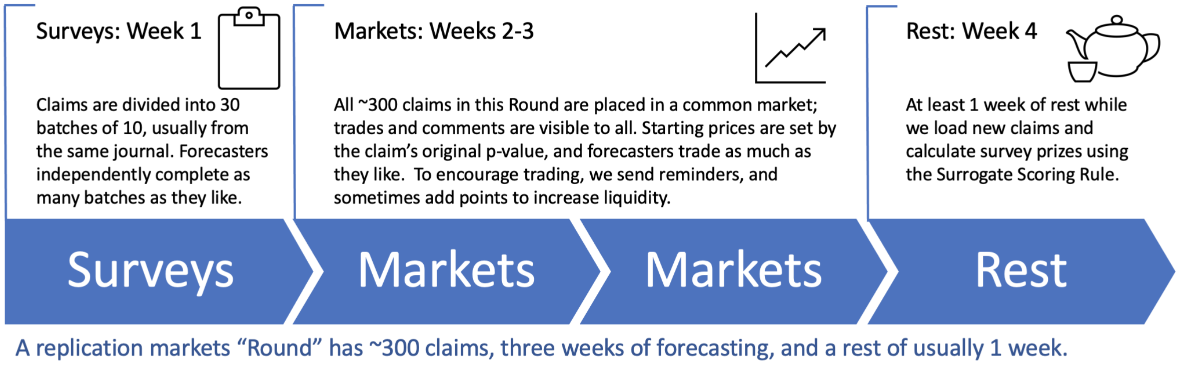
Replication markets

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 (Camerer et al., 2016, 2018; Dreber et al., 2015; Ebersole et al., 2020; Forsell et al., 2018; Klein et al., 2018; Gordon et al., 2021). This approach and evidence build on the well-established ability of markets to aggregate information efficiently (Arrow et al., 2008; Malkiel & Fama, 1970; Plott et al., 2003; Plott & Sunder, 1988; Radner, 1979). In a number of contexts (K.-Y. Chen et al., 2003; Forsythe et al., 1992, 1999), markets appear to provide better estimates than any one individual can, especially in complex combinatorial prediction markets (Y. Chen & Pennock, 2010) where individuals make systematic errors (Wang et al., 2011).

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 (~⅔), and one part dedicated to the survey (~⅓). 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 (Liu et al., 2020) 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 (Pfeiffer et al., 2020) 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 (Gordon et al., 2020).

Figure 4. Overview of the Replication Markets workflow.



Surveys: Week 1

Claims are divided into 30 batches of 10, usually from the same journal. Forecasters independently complete as many batches as they like.

Markets: Weeks 2-3

All ~300 claims in this Round are placed in a common market; trades and comments are visible to all. Starting prices are set by the claim’s original p-value, and forecasters trade as much as they like. To encourage trading, we send reminders, and sometimes add points to increase liquidity.

Rest: Week 4

At least 1 week of rest while we load new claims and calculate survey prizes using the Surrogate Scoring Rule.

A replication markets “Round” has ~300 claims, three weeks of forecasting, and a rest of usually 1 week.

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.

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., Open Science Collaboration, 2015).

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 (Du et al., 2021, 2018; F. Yang et al., 2018). 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 (Nakshatri et al., 2021). The team’s feature extraction and representation (FEXRep) framework extracts bibliometric, bibliographic, statistical and semantic features from scientific papers (Lanka et al., 2021; Modukuri et al., 2021; Wu et al., 2021, 2020). 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.

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 (K.-Y. Chen et al., 2003; Dasigi et al., 2017; Huber & Carenini, 2019). 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.

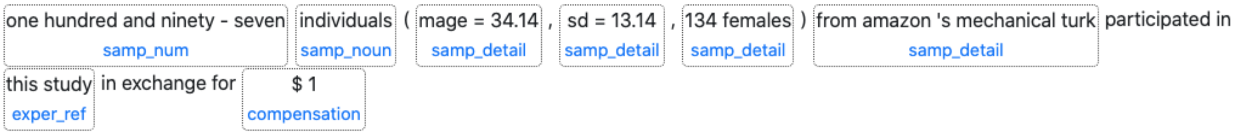
Table 4: Discourse classes used in semantic parsing for the A+ method (TwoSix)

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.

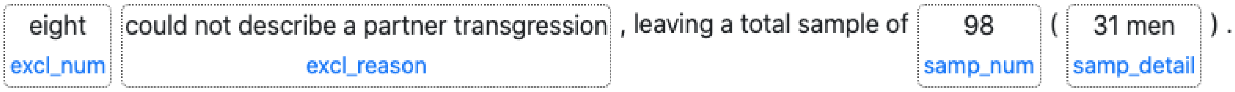
Our model next processes the resulting classified spans -- as shown in Figures 5, 6, and 7-- to opportunistically extract domain-specific numerical and Boolean features. For example, the sample count and exclusion count are both expected to be integers, so it attempts to coerce "one hundred and ninety - 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, R^2, d, F-tests, T-tests, mean, median, standard deviation, confidence intervals, odds ratios, and non-significance.

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



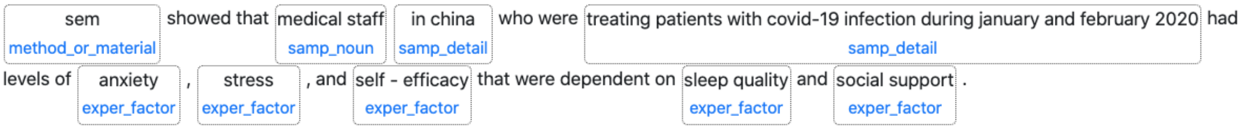
one hundred and ninety - seven individuals (mage = 34.14 , sd = 13.14 , 134 females) from amazon 's mechanical turk participated in this study in exchange for $ 1

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



eight could not describe a partner transgression , leaving a total sample of 98 (31 men) .

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



sem showed that medical staff in china who were treating patients with covid-19 infection during january and february 2020 had levels of anxiety , stress , and self - efficacy that were dependent on sleep quality and social support .

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.

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.

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 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 (Nosek & Errington, 2020).

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 (Open Science Collaboration, 2015) -- 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.

Evaluating Expert and Machine Success

There is no definitive criterion for deciding whether a finding is successfully replicated or reproduced (Nosek et al., 2021), 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” (Yaniv et al., 1991) – 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 (Gibbons & Fielden, 1993), 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 (Pepe, 2003; Steyvers et al., 2014).

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 (Brier, 1950), measures of calibration (Arkes et al., 1995), and various “confusion matrix”-style measures of classification performance (e.g., sensitivity, specificity, proportionate reduction in error vs. base rate; Pepe, 2003). Using metrics based on the p-value to assess replication outcomes have known limitations (Open Science Collaboration, 2015). 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 (Altmejd et al., 2019; Camerer et al., 2018; Dreber et al., 2015; Forsell et al., 2019; Y. Yang et al., 2020).

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 (Gibbons & Fielden, 1993) 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.

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 (Buolamwini & Gebru, 2018; Caliskan et al., 2017; Larson et al., 2016). 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:

1. *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;

2. *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;
3. *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;
4. *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;
5. *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?

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 (Uhlmann et al., 2019), 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.

References

Altmejd, A., Dreber, A., Forsell, E., Huber, J., Imai, T., Johannesson, M., Kirchler, M., Nave, G., & Camerer, C. (2019). Predicting the replicability of social science lab experiments. *PLOS ONE*, *14*(12), e0225826. https://doi.org/10.1371/journal.pone.0225826

Arkes, H. R., Dawson, N. V., Speroff, T., Harrell Jr, F. E., Alzola, C., Phillips, R., Desbiens, N., Oye, R. K., Knaus, W., & Connors Jr, A. F. (1995). The covariance decomposition of the probability score and its use in evaluating prognostic estimates. SUPPORT Investigators. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, *15*(2), 120–131.

Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O., Levmore, S., Litan, R., Milgrom, P., & Nelson, F. D. (2008). The promise of prediction markets. *Science-New York Then Washington-*, *320*(5878), 877.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77–91.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Chang, A. C., & Li, P. (2015). *Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not"* (SSRN Scholarly Paper ID 2669564). Social Science Research Network. https://doi.org/10.2139/ssrn.2669564

Chen, K.-Y., Fine, L. R., & Huberman, B. A. (2003). Predicting the future. *Information Systems Frontiers*, *5*(1), 47–61.

Chen, Y., & Pennock, D. M. (2010). Designing markets for prediction. *AI Magazine*, *31*(4), 42–52.

Cooke, R. M., Marti, D., & Mazzuchi, T. (2021). Expert forecasting with and without uncertainty quantification and weighting: What do the data say? *International Journal of Forecasting*, *37*(1), 378–387. https://doi.org/10.1016/j.ijforecast.2020.06.007

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai van Dongen, N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., … Zhou, X. (2018). Estimating the Reproducibility of Experimental Philosophy. *Review of Philosophy and Psychology*, *12*. https://doi.org/10/gf28qh

Dasigi, P., Burns, G. A., Hovy, E., & de Waard, A. (2017). Experiment segmentation in scientific discourse as clause-level structured prediction using recurrent neural networks. *ArXiv Preprint ArXiv:1702.05398*.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343–15347. https://doi.org/10.1073/pnas.1516179112

Du, M., Liu, N., Song, Q., & Hu, X. (2018). Towards explanation of dnn-based prediction with guided feature inversion. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1358–1367.

Du, M., Liu, N., Yang, F., & Hu, X. (2021). Learning credible DNNs via incorporating prior knowledge and model local explanation. *Knowledge and Information Systems*, *63*(2), 305–332.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., … Nosek, B. A. (2020). Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). An open investigation of the reproducibility of cancer biology research. *ELife*, *3*, e04333. https://doi.org/10.7554/eLife.04333

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., & Dreber, A. (2018). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*, 102117. https://doi.org/10.1016/j.joep.2018.10.009

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

Forsythe, R., Nelson, F., Neumann, G. R., & Wright, J. (1992). Anatomy of an experimental political stock market. *The American Economic Review*, 1142–1161.

Forsythe, R., Rietz, T. A., & Ross, T. W. (1999). Wishes, expectations and actions: A survey on price formation in election stock markets. *Journal of Economic Behavior & Organization*, *39*(1), 83–110.

Fraser, H., Bush, M., Wintle, B., Mody, F., Smith, E., Hanea, A., Gould, E., Hemming, V., Hamilton, D., Rumpff, L., Wilkinson, D. P., Pearson, R., Thorn, F. S., Ashton, R., Willcox, A., Gray, C. T., Head, A., Ross, M., Groenewegen, R., … Fidler, F. (2021). *Predicting reliability through structured expert elicitation with repliCATS (Collaborative Assessments for Trustworthy Science)*. MetaArXiv. https://doi.org/10.31222/osf.io/2pczv

Gibbons, J. D., & Fielden, J. D. G. (1993). *Nonparametric measures of association*. Sage.

Gordon, M., Viganola, D., Bishop, M., Chen, Y., Dreber, A., Goldfedder, B., Holzmeister, F., Johannesson, M., Liu, Y., Twardy, C., Wang, J., & Pfeiffer, T. (2020). Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *Royal Society Open Science*, *7*(7), 200566. https://doi.org/10.1098/rsos.200566

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. *ArXiv:2102.00517 [Stat]*. http://arxiv.org/abs/2102.00517

Hanea, A., Wilkinson, D. P., McBride, M., Lyon, A., Ravenzwaaij, D. van, Thorn, F. S., Gray, C. T., Mandel, D. R., Willcox, A., Gould, E., Smith, E., Mody, F., Bush, M., Fidler, F., Fraser, H., & Wintle, B. (2021). *Mathematically aggregating experts' predictions of possible futures*. MetaArXiv. https://doi.org/10.31222/osf.io/rxmh7

Huber, P., & Carenini, G. (2019). Predicting discourse structure using distant supervision from sentiment. *ArXiv Preprint ArXiv:1910.14176*.

Kerr, N. L., & Tindale, R. S. (2004). Group Performance and Decision Making. *Annual Review of Psychology*, *55*(1), 623–655. https://doi.org/10.1146/annurev.psych.55.090902.142009

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating Variation in Replicability: A "Many Labs" Replication Project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, J., Alper, S., Vega, D., Aveyard, M., Axt, J., & Babaloia, M. (2018). Many Labs 2: Investigating Variation in Replicability Across Sample and Setting. *Advances in Methods and Practice in Psychological Science*.

Lanka, S. S. T., Rajtmajer, S. M., & Giles, C. L. (2021). *Extraction and evaluation of statistical information from social and behavioral science papers.* Workshop on Scientific Knowledge (Sci-K) at The Web Conference.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)*, *9*(1).

Liu, Y., Wang, J., & Chen, Y. (2020). Surrogate Scoring Rules. *Proceedings of the 21st ACM*

Conference on Economics and Computation, 853–871. https://doi.org/10.1145/3391403.3399488

Malkiel, B. G., & Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417.

McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2008). Do economics journal archives promote replicable research? *Canadian Journal of Economics/Revue Canadienne d'économique*, *41*(4), 1406–1420. https://doi.org/10.1111/j.1540-5982.2008.00509.x

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, *21*(1), 1–14. https://doi.org/10.1037/xap0000040

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015). Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspectives on Psychological Science*, *10*(3), 267–281. https://doi.org/10.1177/1745691615577794

Modukuri, S. A., Rajtmajer, S., Squicciarini, A. C., Wu, J., & Giles, C. L. (2021). *Understanding and Predicting Retractions of Published Work*.

Nakshatri, N., Menon, A., Giles, C. L., Rajtmajer, S., & Griffin, C. (2021). Design and Analysis of a Synthetic Prediction Market using Dynamic Convex Sets. *ArXiv Preprint ArXiv:2101.01787*.

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2021). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*.

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

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

Pearson, R., Fraser, H., Bush, M., Mody, F., Widjaja, I., Head, A., Wilkinson, D. P., Sinnott, R., Wintle, B., & Burgman, M. (2021). Eliciting group judgements about replicability: A technical implementation of the IDEA Protocol. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 461.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Medicine.

Pfeiffer, T., Chen, Y., Viganola, D., Bishop, M., Dreber, A., Johannesson, M., & Twardy, C. R. (2020). Prereg Replication Markets Rounds 1-10 (Version 7). *Open Science Framework*. osf.io/svg3x

Plott, C. R., & Sunder, S. (1988). Rational expectations and the aggregation of diverse information in laboratory security markets. *Econometrica: Journal of the Econometric Society*, 1085–1118.

Plott, C. R., Wit, J., & Yang, W. C. (2003). Parimutuel betting markets as information aggregation devices: Experimental results. *Economic Theory*, *22*(2), 311–351.

Radner, R. (1979). Rational expectations equilibrium: Generic existence and the information revealed by prices. *Econometrica: Journal of the Econometric Society*, 655–678.

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, *30*(2), 344–356. https://doi.org/10.1016/j.ijforecast.2013.09.009

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A.

(2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 1–7. https://doi.org/10.1038/s41562-020-0912-z

Steyvers, M., Wallsten, T. S., Merkle, E. C., & Turner, B. M. (2014). Evaluating probabilistic forecasts with Bayesian signal detection models. *Risk Analysis*, *34*(3), 435–452.

Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., McCarthy, R. J., Riegelman, A., Silberzahn, R., & Nosek, B. A. (2019). Scientific utopia III: Crowdsourcing science. *Perspectives on Psychological Science*, *14*(5), 711–733.

Wang, G., Kulkarni, S. R., Poor, H. V., & Osherson, D. N. (2011). Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Analysis*, *8*(2), 128–144.

Wintle, B., Mody, F., Smith, E. T., Hanea, A., Wilkinson, D. P., Hemming, V., … Fidler, F. (2021, May 4). Predicting and reasoning about replicability using structured groups. https://doi.org/10.31222/osf.io/vtpmb

Wood, B. D. K., Müller, R., & Brown, A. N. (2018). Push button replication: Is impact evaluation evidence for international development verifiable? *PLOS ONE*, *13*(12), e0209416. https://doi.org/10.1371/journal.pone.0209416

Wu, J., Nivargi, R., Lanka, S. S. T., Menon, A. M., Modukuri, S. A., Nakshatri, N., Wei, X., Wang, Z., Caverlee, J., & Rajtmajer, S. M. (2021). Predicting the Reproducibility of Social and Behavioral Science Papers Using Supervised Learning Models. *ArXiv Preprint ArXiv:2104.04580*.

Wu, J., Wang, P., Wei, X., Rajtmajer, S., Giles, C. L., & Griffin, C. (2020). Acknowledgement Entity Recognition in CORD-19 Papers. *Proceedings of the First Workshop on Scholarly Document Processing*, 10–19.

Yang, F., Liu, N., Wang, S., & Hu, X. (2018). Towards interpretation of recommender systems with sorted explanation paths. *2018 IEEE International Conference on Data Mining (ICDM)*, 667–676.

Yang, Y., Youyou, W., & Uzzi, B. (2020). Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences*, *117*(20), 10762–10768. https://doi.org/10.1073/pnas.1909046117

Yaniv, I., Yates, J. F., & Smith, J. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, *110*(3), 611.