Predicting Reliability Through Structured Expert Elicitation with repliCATS (Collaborative Assessments for Trustworthy Science)

Fraser, Hannah; Bush, Martin; Wintle, Bonnie; Mody, Fallon; Smith, Eden; Hanea, Anca; Gould, Elliot; Hemming, Victoria; Hamilton, Dan; Rumpff, Libby

Published in:
PLoS ONE

Predicting reliability through structured expert elicitation with repliCATS (Collaborative Assessments for Trustworthy Science)

Hannah Fraser[1αβγδζ] , Martin Bush[1αβγδζ *], Bonnie C. Wintle[1βγδζ] , Fallon Mody[1βγδζ] , Eden Smith[1βγδζ] , Anca Hanea[1,2βγδζ] , Elliot Gould[1,2βγδεζ], Victoria Hemming[1,3βγδ], Daniel G. Hamilton[1βδ], Libby Rumpff[1,2,4βγδ], David P. Wilkinson[1,4δεζ] , Ross Pearson[1γδε], Felix Singleton Thorn[1γεζ] , Raquel Ashton[1δ], Aaron Willcox[1δε], Charles T. Gray[1,5δε], Andrew Head[1δ], Melissa Ross[1βδ], Rebecca Groenewegen[1,4ε], Alexandru Marcoci[6βγδ], Ans Vercammen[7βγδζ], Tim Parker[8βγδ], Rink Hoekstra[9βγδ], Shinichi Nakagawa[10βγδ], David R. Mandel[11β], Don van Ravenzwaaij[9βδζ] , Marissa McBride[7δζ] , Richard O. Sinnott[12βε], Peter Vesk[1,4γ], Mark Burgman[7βγδ], Fiona Fidler[1βγδζ]

1 MetaMelb Lab, University of Melbourne, Melbourne, Victoria, Australia
2 Centre of Excellence for Biosecurity Risk Analysis, University of Melbourne, Melbourne, Victoria, Australia
3 Martin Conservation Decisions Lab, Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, Canada
4 Quantitative & Applied Ecology Group, University of Melbourne, Melbourne, Victoria, Australia
5 School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, UK
6 Philosophy Department, University of North Carolina, Chapel Hill, North Carolina, USA
7 Centre for Environmental Policy, Imperial College London, London, UK
8 Department of Biology, Whitman College, Walla Walla, Washington, USA
9 Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands
10 School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales, Australia
11 Department of Psychology, York University, Toronto, Ontario, Canada
12 Melbourne eResearch Group, University of Melbourne, Melbourne, Victoria, Australia

α These are joint first authors of this paper.
β These authors contributed to the writing.
γ These authors contributed to conceptualization and methodology.
δ These authors contributed to investigation.
ϵ These authors contributed to data curation.
ζ These authors contributed to analysis.
Fiona Fidler supervised the project.

* Corresponding author: martin.bush@unimelb.edu.au

Abstract

Replication is a hallmark of scientific research. As replications of individual studies are resource intensive, techniques for predicting the replicability are required. We introduce a new technique to evaluating replicability, the repliCATS (Collaborative Assessments for Trustworthy Science) process, a structured expert elicitation approach based on the IDEA protocol. The repliCATS process is delivered through an underpinning online platform and applied to the evaluation of research claims in social and behavioural sciences. This process can be deployed for both rapid assessment of small numbers of claims, and assessment of high volumes of claims over an extended period. Pilot data suggests that the accuracy of the repliCATS process meets or exceeds that of other techniques used to predict replicability. An important advantage of the repliCATS process is that it collects qualitative data that has the potential to assist with problems like understanding the limits of generalizability of scientific claims. The repliCATS process has potential applications in alternative peer review and in the allocation of effort for replication studies.

1. Introduction

Scientific claims should be held to a strong standard. The strongest claims will be *reliable*, in that multiple observers examining the same data should agree on the facts/results, *replicable*, in that repetitions of the methods and procedures should produce the same facts/results, and *generalizable*, in that claims should extend beyond a single dataset or process. Evaluating the strength of scientific claims on the basis of these criteria, however, is not straightforward.

Replication studies are a key technique for assessing the strength of evidence in particular studies and claims made in resultant research papers. They contribute to the progressive development of robust knowledge with respect to both inferences from empirical data and deductions made from empirical evidence; policy-making and public trust both draw on such developments. Interestingly, several large-scale replication projects in the social sciences and other disciplines have shown that many published claims do not replicate [1–4]. These failures to replicate may indicate false positives in the original study, or they may have other explanations, such as differences in statistical power between the original and the replication, or unknown moderators in either.

Barriers such as logistics, opportunity, expense and career incentive all militate against replicating published studies. Datasets like national censuses are not feasibly recreated. Historical circumstances are impossible to study again. Even where the contextual factors are more favourable, the expense of full replication studies requires analysis of the potential benefits against costs [5]. Therefore, analytical techniques for a prognostic assessment of the reliability, replicability and generalizability of research claims without attempting full replications are of substantial value because they can provide similar benefits at much lower cost. They can also inform decisions about where to direct scarce resources for such full replications.

This paper outlines an expert elicitation protocol designed to accurately predict the replicability of a large volume of claims across the social and behavioural sciences. We also describe results from a pilot study that suggests this is a fruitful approach for predicting replicability. We discuss how this technique has the potential to provide information about aspects of the credibility of scientific claims beyond replicability, such as the generalizability of claims.

The elicitation procedure in this paper is based on the IDEA protocol [6, 7], a modified Delphi process, with experts working in small groups to provide quantitative predictions of the probability of successful replication. The elicitation also gathers

qualitative data to investigate the reasoning behind predictions of replicability. This qualitative data provides insights into participants' judgements about the credibility signals of claims beyond replicability.

The IDEA protocol was applied to assessing the replicability of research claims in the social and behavioural sciences by the repliCATS (Collaborative Assessments for Trustworthy Science) project as a component of the Systematizing Confidence in Open Research and Evidence (SCORE) program funded by the US Defence Advanced Research Projects Agency. The overall goal of the SCORE program is to create automated tools for forecasting the replicability of research claims made within the social and behavioural science literature. In Phase 1 of the SCORE program, we elicited expert assessments of replicability for 3000 research claims. These assessments formed a benchmark for the comparison of the performance of automated tools developed by other teams within the SCORE program. In Phase 2 of the SCORE program we will elicit assessments of replicability for another 3000 research claims while expanding the elicitation to include additional credibility signals.

2. Previous approaches to predicting replicability

Several previous studies have attempted to predict the outcomes of replication studies. These have typically run alongside large-scale replication projects and used replication outcomes as the ground truth to test prediction accuracy. The two main techniques that have been used for human-derived predictions are surveys and prediction markets [2, 8]. In the former, experts give independent judgements which are aggregated into quantitative predictions of replicability. In the latter, participants trade contracts that give a small payoff if and only if the study is replicated and prices are used to derive the likelihood of replicability. Some other studies have fully automated predictions of replicability, such as machine learning techniques [9, 10] although such techniques are not considered further in this paper.

There are multiple ways of measuring the success of human predictions. The most straightforward is 'classification accuracy'. For this we treat predictions >50% as predictions of replication success and <50% as predictions of replication failure. Classification accuracy is the percentage of predictions that were correct (i.e. on the right side of 50%), excluding predictions of 50%. The classification accuracy from previous prediction studies has ranged between 61% and 86%. The lower limit was reported by Camerer et al. [1] for both surveys and prediction markets on the replicability of 18 laboratory experiments in economics. The higher limit was from Camerer et al. [2] for surveys and prediction markets on predicting the replicability of 21 social science experimental studies published in Nature and Science between 2010 and 2015. In both studies, surveys and prediction markets performed well. Other studies have reported small but noteworthy differences. For the surveys and markets running alongside the Many Labs 2 project [3], prediction markets had a 75% classification accuracy while pre-market surveys reached 67%. There is also evidence that for some kinds of social science research claims non-experts are able to make fairly accurate predictions for the replicability of research claims [11].

3. Design of the repliCATS process

3.1 The IDEA protocol

The repliCATS project introduces a new approach to predicting the replicability of research claims. This is neither a prediction market, nor a simple one-off survey. The IDEA protocol, which forms the basis of the repliCATS approach, involves four steps,

represented in the acronym IDEA: 'Investigate', 'Discuss', 'Estimate' and 'Aggregate' (Fig 1). Each individual is provided a scientific claim and the original research paper to read, and provide an estimate of whether or not the claim will replicate (*Investigate*). They then see the group's judgements and reasoning, and can interrogate these (*Discuss*). Following this, each individual provides a second private assessment (*Estimate*). A mathematical aggregation of the individual estimates is taken as the final assessment (*Aggregate*).

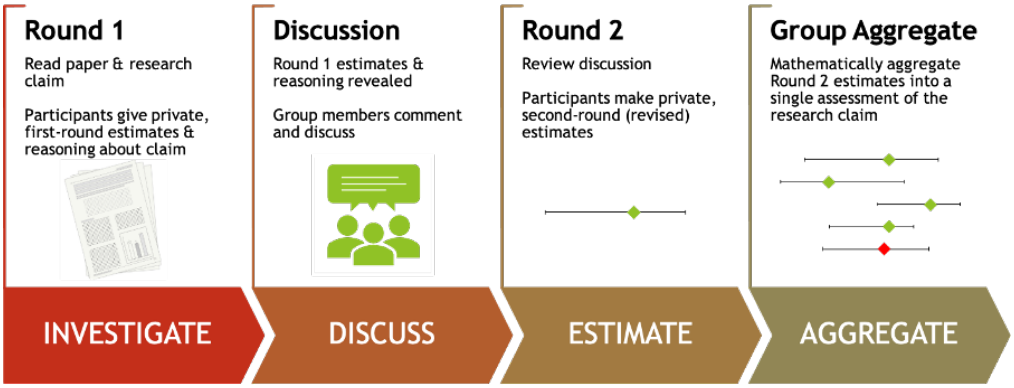

Fig 1. Overview of the IDEA protocol, as adopted in the repliCATS project.

3.2 Research goals

In Phase 1 of the SCORE program, the repliCATS project had three main research goals: *accuracy*, *scalability* and *insight*. That is: to match or improve previous quantitative predictions of replicability; to develop an efficient and scalable process able to be applied to large volumes of claims; and to collect qualitative data that contribute to evaluating credibility signals beyond replicability. The IDEA protocol was applied to each of these goals in a range of ways.

An essential goal of the repliCATS project is to improve upon the already good *accuracy* of current techniques for predicting replicability. Fine tuning expert accuracy in forecasting replicability is a non-trivial challenge. The primary ways in which the IDEA protocol is expected to meet the exacting demands of this task are by harnessing the wisdom of the crowd and diverse ideas, sharing information between participants to resolve misunderstandings and promote further counterfactual thinking, and by taking an aggregate of the judgement. These help to improve accuracy, and reduce overconfidence by reducing biases such as anchoring, groupthink, and confirmation bias.

The IDEA protocol aims to improve accuracy through controlling groupthink by aggregating group assessments *mathematically* and not *behaviourally* [12–15]. That is, group members are not forced to agree on a single final judgement that reflects the whole group. Mathematical aggregation can be more complex than taking the arithmetic mean. Several aggregation techniques have been pre-registered by repliCATS

(https://osf.io/m6gdp/). Described in detail in Hanea et al (in review), these fall into three main groupings: 1) linear combinations of best estimates, transformed best estimates [16] and distributions [17]; 2) Bayesian approaches, one of which incorporates characteristics of a claim directly from the paper, such as sample size and effect size (Gould et al, in prep); and 3) linear combinations of best estimates, mainly weighted by potential proxies for good forecasting performance, such as demonstrated breadth of reasoning, engagement in the task, openness to changing opinion, informativeness of judgements, and prior knowledge (inspired by Mellers et al. [18, 19]).

There are other hypothesised benefits of the IDEA protocol in terms of accuracy. Groupthink is also controlled by recruiting groups that are, ideally, as diverse as possible (Page, 2008). Work in structured elicitation has described how the sharing of information can improve the accuracy of group judgements [20]. The IDEA protocol implemented by the repliCATS project implements this feature, in contrast with survey-based methods of prediction, which generally do not allow for the sharing of information between participants. The IDEA protocol also reduces overconfidence in individual judgements through the use of three-point elicitations, that is asking participants to provide lower and upper bounds for their assessment, as well as their best estimate [21, 22]. This technique [23, 24] is thought to encourage information sampling [23] and prompt participants to consider counter-arguments [25]. (However, see also [26] for counter-evidence regarding this sequencing.)

In addition to accuracy, the repliCATS project aims to provide a *scalable* process. The IDEA protocol is typically implemented through group sizes between four and seven, with perhaps one or two groups of experts per problem, each working under the guidance of a facilitator. This process allows for a rapid evaluation of claims, in contrast with prediction markets that rely on many participants engaged in multiple trades. At the same time, the repliCATS online platform (see [27] section 3.3) allowed an implementation of the IDEA protocol for many experts addressing many problems, with the capacity for the assessment of 3000 claims in 18 months.

Finally, the repliCATS project will generate valuable qualitative data to provide *insight* on issues beyond the direct replicability of specific evidentiary claims. Such problems include: identifying the precise areas of concern for the replicability of a given claim; understanding the limits of generalizability and hence potential applicability of a claim for a research end-user, and assessing the quality of the operationalization of a given research study design. For example, a research claim may be highly replicable and yet offer a poorly operationalized test of the target hypothesis [28]. In such a case, knowing only the predicted replicability of the claim says little about the status of the overall hypothesis. Similarly, even if a claim is well-operationalized, understanding how applicable a claim is for research end-users means understanding its limits of generalizability. This aspect of the repliCATS project, as an implementation of the IDEA protocol, is arguably the most critical point of difference from previous approaches to predicting replicability. Although both surveys and prediction markets can be adapted to collect this kind of data, the advantage of the repliCATS process described here is that such data is produced directly through the elicitation itself, resulting in more straightforward and richer data generation, collection and subsequent analysis.

3.3 The repliCATS platform

To implement our approach we developed a cloud-based, multi-user software platform ('the repliCATS platform') that supports both synchronous face-to-face workshops and asynchronous remote group elicitations [27]. A full description of the elicitation appears in the S1 Appendix. Fig 2 provides a snapshot of the technical operationalization of this elicitation as supported in the repliCATS platform. In particular, Fig 2 shows an

example of aggregated group judgements and reasoning from round 1, as shown to participants in round 2 prior to submitting their final judgements and reasoning.

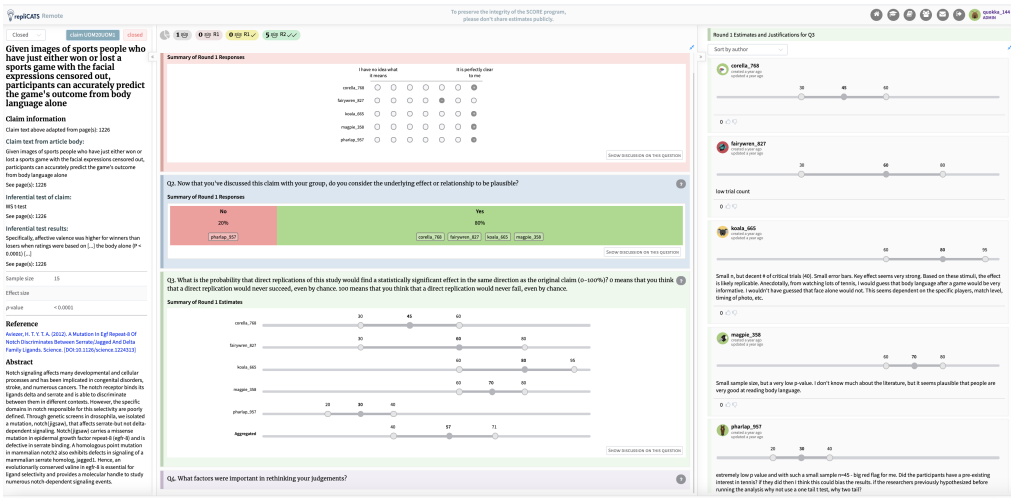

Fig 2. Feedback provided to participants about the group responses to the question about the probability of finding the same result in a replication study.

The repliCATS platform also incorporates several technical features designed to improve participants' confidence in completing their assessments, as well as support or enhance the elicitation process. These features include ready access to general decision support materials and a glossary; tooltips for each question to provide additional guidance on answering each question; additional interactive discussion elements such as the ability to upvote, downvote and have threaded comments; and gamification in the form of participant badges. The badges were designed to reward participation as well as to reward behaviours considered to be beneficial for improving participants' and group judgements. For example, participants were awarded badges for seeking out decision support materials, interacting with group members' reasoning, and consistently submitting their reasoning while evaluating claims.

4. Current applications of the repliCATS process

The primary current application of the repliCATS process has been through the SCORE program. Alongside this, the platform has been used in a number of experiments and classroom teaching contexts as briefly described below.

Phase 1 of the SCORE project ran from February 2019 to 30 November 2020. In this phase, the repliCATS project was provided with 3000 social and behavioural science claims for evaluation by another team involved in the SCORE program. This independent team was also responsible for conducting replication studies for a subset of the 3000 claims. The claims were drawn from 62 peer-reviewed journals across eight social and behavioural science disciplines: business research, criminology, economics, education, political science, psychology, and sociology. Full details of these journals are provided in the S2 Appendix.

Expert predictions of replicability were gathered through a combination of face-to-face and asynchronous workshops, as well as fully remote elicitations. One third of the claims, as well as the data for the pilot study presented below, were elicited in three large workshops held alongside relevant conferences. Two of the workshops were held alongside the Society for Improvement of Psychological Science (SIPS) conference

in July 2019 (Netherlands) and May 2020 (virtual workshop); and the third was alongside the inaugural Association for Interdisciplinary Metaresearch & Open Science (AIMOS) conference in November 2019 (Melbourne). In workshops, participants were divided into small groups with a facilitator and assessed claims synchronously. Different workshop groups took differing amounts of time for each claim but on average assessments were completed in under 30 minutes.

4.1 Community and Participation

In Phase 1, 759 participants were registered on the repliCATS platform. Of these, 550 users assessed one or more of the 3000 claims. Because of confidentiality requirements, not all online participants could be matched with individual user identities, nor could all student participants be matched to demographic profiles. Participant activities were approved by the Human Research Ethics Committee of The University of Melbourne, ethics ID number 1853445. For the teaching use of the repliCATS platform described in section 6.2, additional ethics clearance was obtained from the University of North Carolina.

4.2 Pilot study

To validate and test the process, we conducted a pilot workshop (Wintle et. al., in prep.) alongside the SIPS conference in 2019. This experiment was conducted independently of the SCORE program. A pre-registration for this experiment can be seen at https://osf.io/e8dh7/. Five groups of five participants separately assessed the replicability of 25 claims drawn from previous replication projects. When each group's data was aggregated, the pilot groups achieved a classification accuracy of 84% (and an *Area Under the Curve* measure of 0.94). This indicates that the repliCATS technique performed comparably or better than existing methods for predicting replicability. However, this value should be interpreted with caution. We expected that the accuracy of the participants might be higher for the pilot claims than for the 3000 SCORE claims. The majority of articles in the pilot study were from psychology papers. As most participants at the SIPS conference were psychology researchers, they may have known more about the signals or replicability in psychology than in other disciplines. It was also expected that in some instances participants were familiar with the results of replications for the claims they evaluated, as these results were publicly available prior to SIPS 2019.

4.3 Qualitative data

The repliCATS process generates data that describes participants' reasoning about the claims being evaluated. This typically includes justifications for the assessment of the target question about replicability, as well as judgements about the papers' importance, clarity and logical structure. While such reasoning data has been a part of previous applications of the IDEA protocol, it has not been the focus of previous studies. The repliCATS project thus extends the IDEA protocol through a structured qualitative analysis of a substantial corpus.

This data has the potential to address a number of questions that are of interest for replication studies, for research end-users, and more broadly in metaresearch and the philosophy of the sciences. Some of these target problems, like understanding the quality of operationalization of a given research design, or the limits of generalizability of a particular claim, were described in Section 3.2 above. Other issues include identification of particular strengths or weaknesses in a given research claim, with implications for how a claim might be further investigated. Participants' justifications might indicate

weaknesses with a particular kind of measurement technique, suggesting that further work should be focused on that aspect of experimental design. Alternatively, they might indicate a very high level of prior plausibility for the claim, in which case further work on the theoretical background of the claim might be more fruitful.

The analysis of this qualitative data is described in Bush et al. (in prep). In brief, a subset of the data (776 of the set of 3000 claims in Phase 1 of the SCORE program, which comprised 13901 unique justifications from a total dataset of 46408 justifications) was coded by a team of five analysts against a shared codebook, with each justification being coded by at least two coders. Several principles were applied in developing the codebook for analysis. Inclusion and exclusion criteria were developed for codes most relevant to the research questions. In particular, this included both direct markers of replicability (e.g. statistical details) and proxy markers of replicability (e.g. clarity of writing). The total size of the codebook was limited to ensure its usability. The codebook was developed iteratively, with discussion among coders between each round of development to provide a form of contextualised content analysis. In addition, the inter-coder-reliability (ICR) for each code was assessed at various points in this process. The ICR results were used for reviewing the convergence of textual interpretation between analysts, and after coding, mixed-method aggregation techniques only utilised those codes that had met a pre-registered ICR.

5. Challenges for the repliCATS process

The advantages of the IDEA protocol described above also come with challenges. Some of these challenges were due to the nature of the elicitation, some were based on research goals of the repliCATS project, and some were based on the broader SCORE program.

5.1 Challenges with elicitation

A problem for any research involving human subjects is elicitation burden – both the quantity and quality of information provided by experts are thought to decline the longer the elicitation process takes, similar to participant fatigue in surveys and experiments [29]. There is thus a trade-off between the extent of information desired for research purposes and the number and richness of questions asked. As noted, the average time taken for a group to assess a claim in repliCATS workshops was just under 30 minutes. Such a bounded time of participation was also necessary because the majority of research participants were volunteers who were unpaid, or only minimally compensated for their time. Although aspects of the elicitation are believed to be intrinsically rewarding - by design - this is another reason for minimizing the burden on participants.

In order to address the elicitation burden, the repliCATS platform was designed to be user-friendly, and online workshops were flexible in terms of participant schedules. However, the trade-off for such elicitations is that it reduced the capacity for participants to engage in discussion, which is a key aspect of the IDEA protocol. Participants in widely-spaced time zones could not have synchronous discussions, and could only exchange comments over a period of days. Future online workshops using the repliCATS process will aim to group participants, as best as possible into compatible time zones, although other logistical constraints do not always make this possible. (By contrast, face-to-face workshops allow for full discussions, at the cost of committing participants to a set schedule and physical co-location.)

A related difficulty is that research participants might have multiple interpretations of key terms for the elicitation. Indeed, there can be considerable conceptual slippage around many of these terms, such as conflicting taxonomies attempting to define

replication practices. While the discussion phase of the IDEA protocol can be used to work on resolving such ambiguities, it can be difficult or impossible to completely eliminate such differences in interpretation. This difficulty was addressed by encouraging participants to describe their interpretations through textual responses, allowing both team members and qualitative analysts to understand these ambiguities better. The repliCATS platform also provides the definitions for key terms such as 'successful replication' where specific definitions are used within the SCORE program. The provision of such information also involves trade-offs between definitional specificity and richness of response.

Another challenge is that participants may hold opinions about aspects of a study that could bias their judgements about replicability. One such issue relates to the perceived importance of a study, whilst another relates to stylistic features of how the paper is written. While both of these may be proxies for the replicability of a study in some contexts, they need not be. Even more directly relevant markers, like the quality of the experimental design, need not suggest non-replicability. For example, a poorly operationalized study where the dependent variable is auto-correlated with the independent variable will, in fact, be highly replicable if you repeat the study exactly. For these reasons, the elicitation was designed to separate out judgements about replicability while allowing participants to explicitly express their opinions about matters like the clarity, plausibility and importance of the study. This is done through two questions prior to the question about replicability, and one question afterwards. Figure 2 shows an example of one of these questions, about the comprehensibility of the research claim, answered on a 7-point scale.

5.2 Challenges with recruitment

Ideally, for most forecasting problems, IDEA groups are deliberately constructed to be diverse. People differ in the way they perceive and analyze problems, as well as the knowledge they bring to bear on them. Access to a greater variety of information and analyses may improve individual judgements. The accuracy of participants' answers to quantitative and probabilistic judgements has been shown to improve between peoples' independent initial judgements and the final judgements they submit through the IDEA protocol [6, 7]. Offset against this is that domain-specific knowledge is clearly relevant to understanding the details of research claims and thus being able to assess their replicability. The balance between diversity and domain knowledge in good assessments of replicability is poorly understood. This problem warrants further research. The repliCATS recruitment strategy meant that we were not able fully to control for or consistently recruit diversity in groups. Where participants are allowed to self-select claims to assess, such diversity is even harder to achieve, even if the overall participant pool contains a large amount of diversity.

5.3 Challenges posed by the SCORE program

The main challenge presented by the SCORE program was the need to evaluate 3000 research claims in approximately 12 months (allowing for platform development time at the start of Phase 1 and analysis of results at the end). Preferably, multiple groups would have assessed each research claim and assessments compared and aggregated across groups, as was done with the pilot study (Section 4.1 above). The high volume of claims involved, however, precluded this as a feasible approach.

6. Future applications of the repliCATS approach

The repliCATS process has intended applications beyond its use in the SCORE program. Work on some of these potential applications has begun as described below.

6.1 Capacity building in peer review

There are few peer review training opportunities available to researchers [30] and even fewer that have clearly demonstrated evidence of success [31, 32]. As a consequence, many early career researchers report that they frequently learn how to review in passive and fragmentary ways, such as performing joint reviews with their advisors and senior colleagues, e.g. via participation in journal clubs or from studying reviews of their own submissions [33].

Participants in the repliCATS project noted the benefits of the feedback and calibration in the repliCATS protocol: "I got a lot of exposure to a variety of research designs and approaches (including some fun and interesting theories!) and was afforded [the] opportunity to practice evaluating evidence. In practicing evidence evaluation, I feel like I sharpened my own critical evaluative skills and learned from the evaluations of others." This type of feedback was particularly common from early career researchers, although more experienced researchers also described the value of the process as professional development in peer review.

This feedback suggests the potential to deploy evaluations of research papers through the repliCATS process as explicit training in both research design and peer review. A pilot application of such student training has already been undertaken at the University of North Carolina, Chapel Hill where the repliCATS process was deployed as an extra credit undergraduate student activity.

6.2 repliCATS as an alternative peer review model

Surprisingly little is known about how reviewers conduct peer review. More than a quarter of a century after the remark was made, it is still the case that "we know surprisingly little about the cognitive aspects of what a reviewer does when he or she assesses a study" [34]. Nor are journal instructions to reviewers especially informative here with criteria for acceptance usually expressed in general terms, such as "how do you rate the quality of the work?".

Initiatives such as Transparency and Openness Promotion (TOP) guidelines offer guidance for authors and reviewers at signatory journals [35] (see also https://www.cos.io/initiatives/top-guidelines). These are important for ensuring completeness of scientific reporting. The repliCATS project is not an attempt at competing guidelines. Its goal is not to provide checklists for reviewers (or authors or editors). Rather, repliCATS reconceptualizes peer review as a process that takes advantage of collective intelligence through an expert deliberation and decision-making process, and thus one to which a structured elicitation and decision protocol is applicable. In phase 1 of the SCORE program, repliCATS focused on providing accurate elicitations of replicability. In phase 2 of the SCORE program, the project scope will be significantly expanded, providing judgements on multiple credibility signals, using the same underlying structured elicitation and decision protocol.

Despite limited implementation among popular journals in the social sciences [36], there are existing models of interactive peer review that encourage increased dialogue between reviewers, such as the one used in the Frontiers journals. However, these typically rely on behavioural consensus, with its attendant disadvantages. In contrast, the repliCATS project has the strong advantage of a predefined end-point, avoiding 'consensus by fatigue'. It is transparent by design, and the underlying IDEA protocol is

directly informed by developments in the expert elicitation, deliberation and decision making literature.

6.3 repliCATS for commissioned review

One particular example of alternative forms of review includes commissioned reviews, such as for papers or research proposals prior to submission. This potential application was suggested by a number of repliCATS participants. Variations on this theme include reviews of a suite of existing published research intended as the basis of a research project or as the evidence base for specific policy, action or management decisions. The former is likely to appeal to early career researchers, for example, at the start of their PhD candidature, and the latter to end users and consumers of research.

6.4 repliCATS as a model for allocating replication effort

Replication of studies is not always possible nor is it always desirable. Some studies cannot be replicated, as described above, due to, for example, their historical nature. Such studies may still have the potential to inform decisions and assessment of their reliability may still be valuable. Nor is replication always ideally suited for assessing the reliability of a claim, even when it is a viable approach. In any case, replications are typically resource-intensive.

There are several approaches to determining how best to allocate scarce resources to replication studies. One suggested approach is to apply the results of prediction markets to this question [8]. Other approaches propose selection of studies for replication based on a tradeoff between the existing strength of evidence for a focal effect and the utility of replicating a given study. In Field et al. [37], a worked example is given on how to combine strength of evidence (quantified with a Bayesian reanalysis of published studies) with theoretical and methodological considerations. Pittelkow et al. [38] follow a similar procedure, applied to studies from clinical psychology. Isager et al. [39] outline a model for deciding on the utility of replicating a study, given known costs, by calculating the value of a claim and the uncertainty about that claim prior to replication. The key variables in this model – costs, value, and uncertainty – remain undefined, with the expectation that each can be specified outside the model (as relevant to a given knowledge domain). This approach to formalizing decisions can help clarify how to justify allocating resources toward specific replication practices.

Complementing the latter approach, the repliCATS process can generate data about specific research claims that can be used as inputs to a formal model. For example, the repliCATS approach is able to provide data on uncertainty, such as the extent of agreement or disagreement within groups about the replicability of a given claim, and the potential uncertainty of individual assessments as seen in interval widths. Textual data in Phase 1 of the SCORE program provided information about the theoretical or practical value of specific claims, and further development of the repliCATS platform in Phase 2 of the SCORE program will develop ways of eliciting this kind of information in a more quantitative form.

7. Conclusion

Assessment of the credibility of scientific papers in general, and predicting the replicability of published research claims in particular is an example of an expert decision-making problem. In such cases the use of a structured protocol has known advantages. The repliCATS platform implements a user-friendly realization of the IDEA protocol for the assessment of published research papers. Experiments have shown that

this platform has the capacity to undertake both rapid review of small numbers of claims while being scalable to a large volume of claims over an extended period. Pilot results suggest that the accuracy of prediction of replicability for research claims by the repliCATS process meets or exceeds previous techniques. Analysis of the full results from Phase 1 of the SCORE program is forthcoming and will shed light on whether repliCATS reliably improves replication prognostics. A particular advantage of the repliCATS process is the collection of rich qualitative data that can be used to address questions beyond those of direct replicability, such as the generalizability of given claims. Future work will expand the assessment of claims made possible through the repliCATS platform to incorporate a broader set of credibility signals, with potential application to support alternative models of peer review.

S1 Appendix. Details of elicitation questions used in the repliCATS process.

Round 1 elicitation questions

1. How well do you understand this claim?
Response format: 7-point scale with free-text box for comments

2. What's your initial reaction: is the underlying effect or relationship plausible?
Response format: Binary Yes/No

3. What is the probability that direct replications of this study would find a statistically significant effect in the same direction as the original claim (0-100%)? 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.
Response format: Three-point elicitation with free-text box for comments

4. Considering the major factors that influenced your thinking in making these judgements, please describe any important aspects that you have not covered above.
Response format: Free-text box for comments

5. Were you involved in the writing, data collection, or analysis of the original study?
Response format: Binary Yes/No

Round 2 elicitation questions

6. Now that you've discussed this claim with your group, how well do you understand it?
Response format: 7-point scale with free-text box for comments

7. Now that you've discussed this claim with your group, do you consider the underlying effect or relationship to be plausible?
Response format: Binary Yes/No

8. What is the probability that direct replications of this study would find a statistically significant effect in the same direction as the original claim (0-100%)? 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.
Response format: Three-point elicitation with free-text box for comments

9. What factors were important in rethinking your judgements?
Response format: Free-text box for comments

S2 Appendix. List of source journals for research claims in phase 1 of the SCORE program.

Criminology
Law and Human Behavior
Criminology

Economics
American Economic Review
Journal of Finance
Quarterly Journal of Economics
Journal of Labor Economics
Journal of Financial Economics
Review of Financial Studies
Econometrica
American Economic Journal: Applied Economics
Experimental Economics

Education
Educational Researcher
American Educational Research Journal
Journal of Educational Psychology
Computers and Education
Learning and Instruction
Contemporary Educational Psychology
Exceptional Children

Health related
Psychological Medicine
Health Psychology
Social Science and Medicine

Management
Journal of Management
Journal of Business Research
Academy of Management Journal
Leadership Quarterly
Organization Science
Management Science

Marketing/Org Behavior
Journal of Marketing
Journal of Consumer Research
Journal of the Academy of Marketing Science
Journal of Marketing Research
Journal of Organizational Behavior
Organizational Behavior and Human Decision Processes

Political Science
American Journal of Political Science
American Political Science Review
Journal of Political Economy
Comparative Political Studies

World Development
Journal of Conflict Resolution
British Journal of Political Science
World Politics
Journal of Experimental Political Science

Psychology
Journal of Personality and Social Psychology
Journal of Applied Psychology
Journal of Consulting and Clinical Psychology
Journal of Experimental Psychology: General Psychological Science
Child Development
Clinical Psychological Science
European Journal of Personality
Cognition
Journal of Experimental Social Psychology
Journal of Environmental Psychology
Evolution and Human Behavior

Public Administration
Journal of Public Administration Research and Theory
Public Administration Review

Sociology
American Sociological Review
Journal of Marriage and Family
Demography
American Journal of Sociology
European Sociological Review
Social Forces

Acknowledgments

This project is sponsored by the Defense Advanced Research Projects Agency (DARPA) under cooperative agreement No.HR001118S0047. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Camerer CF, Dreber A, Forsell E, Ho TH, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. Science; 351(6280):1433–1436. doi:10.1126/science.aaf0918.

2. Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, Johannesson M, et al. Evaluating the rreplicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour; 2(9):637–644. doi:10.1038/s41562-018-0399-z

3. Klein RA, Ratliff KA, Vianello M, Adams RB, Bahník S, Bernstein MJ, et al. Investigating variation in replicability: "Many Labs" Replication project. Social Psychology; 45(3):142–152. doi:10.1027/1864-9335/a000178

4. Open Science Collaboration. Estimating the reproducibility of psychological science. Science; 349(6251):aac4716. doi:10.1126/science.aac4716

5. Coles N, Tiokhin L, Scheel AM, Isager PM, Lakens D. The costs and benefits of replication studies. PsyArXiv [Preprint]. 2018 PsyArXiv c8akj [posted 2018 January 18; revised 2018 July 2; cited 2021 Feb 17]: [7 p.]. Available from: `https://psyarxiv.com/c8akj/` doi:10.31234/osf.io/c8akj

6. Hanea AM, Burgman M, Hemming V. IDEA for uncertainty quantification. In: Dias LC, Morton A, Quigley J, editors. Elicitation: the science and art of structuring judgement. International Series in Operations Research & Management Science. Springer International Publishing; pp. 95–117. doi:10.1007/978-3-319-65052-4_5

7. Hemming V, Burgman MA, Hanea AM, McBride MF, Wintle BC. A practical guide to structured expert elicitation using the IDEA protocol. Methods in Ecology and Evolution; 9(1):169–180. doi:10.1111/2041-210X.12857

8. Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, et al. Using prediction markets to estimate the reproducibility of scientific research. Proceedings of the National Academy of Sciences; 112(50):15343–15347. doi:10.1073/pnas.1516179112

9. Altmejd A, Dreber A, Forsell E, Huber J, Imai T, Johannesson M, et al. Predicting the replicability of social science lab experiments. PLOS ONE; 14(12):e0225826. doi:10.1371/journal.pone.0225826

10. Yang Y, Youyou W, Uzzi B. Estimating the deep replicability of scientific findings using human and artificial intelligence. Proceedings of the National Academy of Sciences; p. 201909046. doi:10.1073/pnas.1909046117

11. Hoogeveen S, Sarafoglou A, Wagenmakers EJ. Laypeople can predict which social science studies replicate. Advances in Methods and Practices in Psychological Science; September 2020:267–285. doi:10.1177/2515245920919667

12. French S. Aggregating expert judgement. Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas; 105(1):181–206. doi:10.1007/s13398-011-0018-6

13. Hanea AM, McBride MF, Burgman MA, Wintle BC. The value of performance weights and discussion in aggregated expert judgments. Risk Analysis: An Official Publication of the Society for Risk Analysis; 38(9):1781–1794. doi:10.1111/risa.12992

14. Hemming V, Hanea AM, Walshe T, Burgman MA. Weighting and aggregating expert ecological judgments. Ecological Applications; 30(4):e02075. doi:10.1002/eap.2075

15. McAndrew T, Wattanachit N, Gibson GC, Reich NG. Aggregating predictions from experts: a review of statistical methods, experiments, and applications. WIREs Computational Statistics; e1514. doi:10.1002/wics.1514

16. Satopää VA, Baron J, Foster DP, Mellers BA, Tetlock PE, Ungar LH. Combining multiple probability predictions using a simple logit model. International Journal of Forecasting; 30(2):344–356. doi:10.1016/j.ijforecast.2013.09.009

17. Cooke RM, Marti D, Mazzuchi T. Expert forecasting with and without uncertainty quantification and weighting: what do the data say?. International Journal of Forecasting; 37(1):378–387. doi:10.1016/j.ijforecast.2020.06.007

18. 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. Journal of Experimental Psychology: Applied; 21(1):1–14. doi:10.1037/xap0000040

19. 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. Perspectives on Psychological Science; 10(3):267–281. doi:10.1177/1745691615577794

20. Kerr NL, Tindale RS. Group performance and decision making. Annual Review of Psychology; 55(1):623–655. doi:10.1146/annurev.psych.55.090902.142009

21. Burgman MA. Trusting judgements: how to get the best out of experts. Cambridge: Cambridge University Press; 2016.

22. Speirs-Bridge A, Fidler F, McBride M, Flander L, Cumming G, Burgman M. Reducing overconfidence in the interval judgments of experts. Risk Analysis; 30(3):512–523. doi:10.1111/j.1539-6924.2009.01337.x

23. Soll JB, Klayman J. Overconfidence in interval estimates. Journal of Experimental Psychology: Learning, Memory, and Cognition; 30(2):299–314. doi:10.1037/0278-7393.30.2.299

24. Teigen KH, Jørgensen M. When 90% confidence intervals are 50% certain: on the credibility of credible intervals. Applied Cognitive Psychology; 19(4):455–475. doi:10.1002/acp.1085

25. Koriat A, Lichtenstein S, Fischhoff B. Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory; 6(2):107–118. doi:10.1037/0278-7393.6.2.107

26. Mandel DR, Collins RN, Risko EF. Effect of confidence interval construction on judgment accuracy: Judgment and Decision Making; 15(5):783–797.

27. 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: Hawaii International Conference on System Science. 2021 [posted 2021 January 5; cited 2021 Feb 17]: [10 p.]. Available from: `http://hdl.handle.net/10125/70666` doi:10.24251/HICSS.2021.055

28. Yarkoni T. The generalizability crisis. PsyArXiv [Preprint]. 2019 PsyArXiv jqw35 [posted 2019 November 22; revised 2020 November 3; cited 2021 Feb 17]: [27 p.]. Available from: `https://psyarxiv.com/jqw35` doi:10.31234/osf.io/jqw35

29. Lavrakas P. Respondent fatigue. In: Encyclopedia of survey research methods. Sage Publications, Inc. Available from: `http://methods.sagepub.com/reference/encyclopedia-of-survey-research-methods/n480.xml` doi:10.4135/9781412963947

30. Patel J. Why training and specialization is needed for peer review: A Case Study of Peer Review for Randomized Controlled Trials. BMC medicine; 12(1):128.

31. Bruce R, Chauvin A, Trinquart L, Ravaud P, Boutron I. Impact of interventions to improve the quality of peer review of biomedical journals: a systematic review and meta-analysis. BMC medicine; 14(1):85. doi:10.1186/s12916-016-0631-5

32. Callaham ML, Tercier J. The relationship of previous training and experience of journal peer reviewers to subsequent review quality. PLoS medicine; 4(1):e40.

33. McDowell GS, Knutsen JD, Graham JM, Oelker SK, Lijek RS. Research culture: co-reviewing and ghostwriting by early-career researchers in the peer review of manuscripts. eLife; 8:e48425. Available from: `https://elifesciences.org/articles/48425` doi:10.7554/eLife.48425

34. Kassirer JP, Campion EW. Peer review: crude and understudied, but indispensable. JAMA; 272(2): 96–97. doi:10.1001/jama.1994.03520020022005

35. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. Science; 348(6242): 1422. doi:10.1126/science.aab2374

36. Hamilton DG, Fraser H, Hoekstra R, Fidler F. Meta-research: journal policies and editors' opinions on peer review. eLife; 9: e62529. Available from: `https://elifesciences.org/articles/62529` doi:10.7554/eLife.62529

37. Field SM, Hoekstra R, Bringmann L, van Ravenzwaaij D. When and why to replicate: as easy as 1, 2, 3?. Collabra: Psychology; 5(1): 46. doi:10.1525/collabra.218

38. Pittelkow M, Hoekstra R, Karsten J, Ravenzwaaij D van. Replication target selection in clinical psychology: a Bayesian and qualitative re-evaluation. Clinical Psychology: Science and Practice; Forthcoming.

39. Isager PM, van Aert RCM, Bahník S, 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. PsyArXiv [Preprint]. 2018 PsyArXiv c8akj [posted 2020 September 2; cited 2021 Feb 17]: [14 p.]. Available from: `https://osf.io/preprints/metaarxiv/2gurz/` doi:10.31222/osf.io/2gurz