# Eliciting group judgements about replicability: a technical implementation of the IDEA Protocol

E.R. Pearson, H. Fraser, M. Bush, F. Mody, I. Widjaja, A. Head, D.P. Wilkinson
B. Wintle, R. Sinnott, P. Vesk, M. Burgman, F. Fidler, University of Melbourne, Australia
*Refer to https://osf.io/qu27k/ for a full list of contributors*
Contact author email: *pearson.r@unimelb.edu.au*

## Abstract

*In recent years there has been increased interest in replicating prior research. One of the biggest challenges to assessing replicability is the cost in resources and time that it takes to repeat studies. Thus there is an impetus to develop rapid elicitation protocols that can, in a practical manner, estimate the likelihood that research findings will successfully replicate.*

*We employ a novel implementation of the IDEA ('Investigate', 'Discuss', 'Estimate' and 'Aggregate) protocol, realised through the repliCATS platform. The repliCATS platform is designed to scalably elicit expert opinion about replicability of social and behavioural science research. The IDEA protocol provides a structured methodology for eliciting judgements and reasoning from groups.*

*This paper describes the repliCATS platform as a multi-user cloud-based software platform featuring (1) a technical implementation of the IDEA protocol for eliciting expert opinion on research replicability, (2) capture of consent and demographic data, (3) on-line training on replication concepts, and (4) exporting of completed judgements. The platform has, to date, evaluated 3432 social and behavioural science research claims from 637 participants.*

*Keywords: IDEA protocol, replicability; expert elicitation; repliCATS.*

## 1. Introduction

Large scale replication projects, such as the Reproducibility Project Psychology 2015 [1] and Begley and Ellis's preclinical replication study [2] have demonstrated low replicability rates, leading to a crisis of confidence in some areas of psychology and preclinical medicine. These problems are not specific to psychology and preclinical medicine. Similar issues exist in economics [3, 4], philosophy [5], social science [6], neuroscience [7] and areas of biology [8].

Outside these large replication projects, the rates of attempted replication studies in the published literature are low. For example, Makel [9] estimates that only 1% of published studies in psychology are attempted replications of previous research. Kelly [10] estimates an even lower proportion of the ecology literature (0.0006%) are attempted replications of previous work. The problem then is both low replication rates, and few attempted replication studies.

Similar arguments have been made for Information Systems (IS) research [11], and Computer Science at large [12, 13]. The issue is particularly relevant for the IS discipline, commonly regarded as a social science [14, 15, 16]. The launch of AIS Transactions on Replication Research (TRR) provides a much needed platform for replicability studies within IS, such as the Systems Replication Project [17] in which 21 replications studies were performed.

While interest in replication studies is accelerating, there are substantial costs to repeating studies as well as significant difficulties in interpreting the results. This creates a strong impetus to develop protocols for reliably estimating the likelihood of the replicability of research findings without undertaking replication studies. Approaches for generating such assessments include machine learning algorithms trained on prior replications, computer-mediated human assessments such as prediction markets, and expert elicitations.

This paper describes the repliCATS (Collaborative Assessment for Trustworthy Science) platform, a multi-user cloud-based rich internet application designed for expert elicitations from group assessments of the replicability of research claims using the IDEA protocol to collect quantitative and qualitative data from small groups of experts. The repliCATS platform operates within the constraints of a broader research program and is designed to optimise participant performance and experience through community resources and notification features. This paper describes design decisions for an IS approach to the research problem within that context.

URI: https://hdl.handle.net/10125/70666
978-0-9981331-4-0

## 2. The research problem: developing an IS to assess replicability

The repliCATS platform described here was developed as part of the SCORE (Systematizing Confidence in Open Research and Evidence) program funded by DARPA (Defence Advanced Research Projects Agency)[1]. The overall goal of the SCORE project is to create automated tools to assign 'confidence scores' to the replicability of research claims within the social and behavioural science literature. The SCORE program consists of three Technical Aspects, comprising multiple independent teams (known as 'performers'). The role of the repliCATS project as part of the second Technical Aspect (TA2) is to elicit human assessments of replicability for 3000 research claims.

This research question reflects the desire for increased confidence in the evidence base provided by scientific literature. Over the last decade, large scale replication studies in psychology and other disciplines have shown that the literature contains a large number of false positive results. Estimates of the replicability of research claims from the social sciences from previous large scale replication projects vary, but are around 50% or less [3, 6, 18]. From the perspective of a research end-user seeking to develop evidence-based programs by relying on results from the published literature, this replicability rate is unacceptably low. Confidence in the evidence base could be increased by undertaking replication studies for research results that are to be relied upon. However, such studies are expensive and time-consuming. Additionally, some experiments may be difficult or impossible to run again. Hence there is a need for techniques that can, with a high degree of reliability, assess the replicability of research claims without performing replication studies.

Two key approaches for generating such assessments of the replicability of prior research without conducting replication studies are as follows. Firstly, machine learning algorithms can analyse scientific papers in order to assign confidence scores to the claims in the paper. Such algorithms can analyse the text of the papers, the quantitative data reported for the research claims or both [19]. Typically, they will be trained with data from replication studies and other prior information. Machine-learning approaches will be developed by teams from the third Technical Aspect (TA3) of the SCORE program. Secondly, computer-mediated human assessments of research claims can build such confidence scores, and previous research is positive about the ability of people to predict research replicability [3, 6, 18]. Prior approaches to assessments of replicability have used prediction markets to generate confidence scores. Another previously used technique for human-generated assessments is the use of survey data. All of these approaches show promise.

Based on experience of team members with expert elicitations, the repliCATS project adopts a novel approach to the generation of confidence scores: aggregating confidence scores from individuals within small groups through the use of a structured elicitation protocol. A novel aspect of this approach is that, alongside quantitative estimates of replicability, rich qualitative data on participant's reasoning about these judgements is collected allowing research on such cognate problems like those of the generalizability of research claims.

## 3. The platform implements the IDEA protocol to assess replicability

Assessing the replicability of a research claim is an example of decision-making with limited information. In such situations, the elicitation of expert judgements is a fruitful technique [20, 21]. Best practice guides researchers away from relying on the judgements of individual experts, to eliciting judgements from multiple experts. There are a range of elicitation techniques, including: structured or unstructured; interactive – with a discussion component - or non-interactive; employing 'behavioural consensus' where the process forces agreement between experts or 'mathematical consensus' where experts are left to disagree with final judgements being aggregated after the fact). The IDEA protocol used by the repliCATS project is a structured, interactive elicitation protocol which employs mathematical aggregation. It has been described in detail elsewhere, but the main features of it that are important for the repliCATS platform are outlined below [22, 20, 21].

IDEA stands for 'Investigate', 'Discuss', 'Estimate' and 'Aggregate ', which indicate the four steps in the protocol's workflow. Described with reference to the research problem here – assessing the replicability of a specific research claim – these are as follows. With the first step, Investigate, group members individually review a specific research claim and provide individual first round estimates for the predefined elicitation questions. In Discuss, participants review each other's judgments and provide comments and feedback through the online platform. These online judgements are anonymised as far as possible to mitigate biases that can occur when the identity of group members is exposed.

[1] https://www.darpa.mil/program/systematizing-confidence-in-open-research-and-evidence

Estimate refers to the second round judgements about the replicability of the research claim provided by participants after discussion and feedback. The final step, Aggregate, takes the individual quantitative estimates provided in the second round and combines them into a single assessment of the replicability of the research claim for the group. The arithmetic mean is a simple example of an aggregate but there are other ways of aggregating individual judgements into a group assessment. Figure 1 outlines the procedure, as operationalised in the repliCATS platform.

## 4. Overview of features and workflows

This paper describes the repliCATS platform as a technical implementation of the IDEA protocol deployed to make assessments about the reliability of research claims in social and behavioural science. We describe design decisions that are required to build a realizable IS within the constraints of the broader SCORE program that is faithful to the IDEA protocol.

A primary feature of IDEA is the aggregation of judgements from individual participants working across two rounds into group assessments. In order to do this within the SCORE program, the repliCATS platform also requires user management and claims management features. Additional aspects of the repliCATS platform are oriented towards optimising the performance and engagement of volunteer research participants to allow the required scalability.

The repliCATS program has run elicitations in both synchronous workshops, where participants intensively assess a pre-defined list of claims, and in a wholly online 'Remote' mode, where virtual teams self-assemble as participants opt-in to the same claim for evaluation. This produces two different workflows for the repliCATS platform which are described here.

### 4.1. A mode for synchronous workshops

The salient features of workshops are that pre-defined groups of participants assess set lists of claims synchronously, typically in a face-to-face setting. One person within the group acts as a facilitator to keep the group to a schedule so that all claims are assessed in the allocated time. Facilitators also prompt discussion if required. Training materials developed for workshop facilitators provide guidance about how to prompt discussion while allowing the range of opinions within the group to be expressed. There are no limitations enforced by the repliCATS platform on the number of members that can participate in a single group. In our own use, group sizes are typically between 4 – 10 participants.

### 4.2. Remote mode for wholly online groups

Remote mode is distinguished from Workshop mode in that participants select which claims they would like to assess, the virtual teams are self-formed, and claims are assessed by participants asynchronously without facilitation. Participants within this workflow can select from a set of claims to which all such participants have access. Virtual teams self-assemble as participants opt-in to assessing the same claim. Once a predefined number of participants complete a second round judgement on the same claim, that claim is closed for access from new participants.

Implementation of Remote mode was essential for the scalability required by the SCORE program - 3000 assessments over 18 months. The implementation of the IDEA protocol in Remote mode, compared with Workshop mode, is better in terms of anonymity of participants but worse in terms of forming diverse teams, as it is to be expected that people of similar expertise will opt-in to similar claims. We note this kind of trade-off is common in practical IS implementations of such protocols.

## 5. User management for different modes

Individuals assessing claims on the repliCATS platform are volunteer research participants and achieving the required scalability requires maximising their participation through platform features such as allowing self-enrollment and control of account details. The platform is required to record informed consent from participants as a baseline requirement for ethical research practice. Basic demographic information is also collected from participants. All this needs to meet confidentiality standards, as described in the platform Architecture section below.

### 5.1. User enrollment across the two modes

The IDEA protocol is based on individual assessments within groups and the repliCATS platform reflects this. Two methods are supported by the platform for participant enrollment: bulk upload via the API and self-enrollment via the web front end, associated with workshop and remote workflows respectively. The user bulk upload proceeds through a Python script that imports a CSV file containing participant details to an API for storage in the repliCATS platform database. Self-enrollment is executed through the web application front end which is available at https://score.eresearch.unimelb.edu.au. One distinction between bulk enrolled participants and self-enrolled participants, is that the former have their accounts

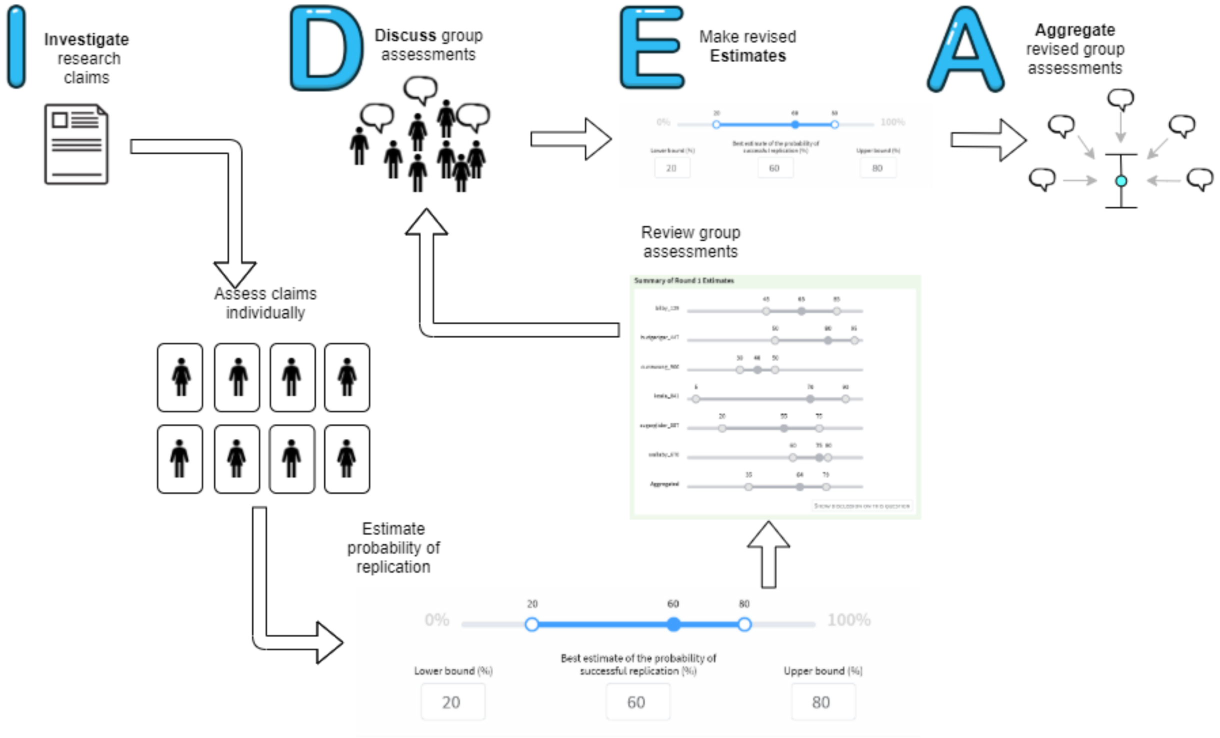

I
Investigate research claims
Assess claims individually
Estimate probability of replication
D
Discuss group assessments
E
Make revised Estimates
A
Aggregate revised group assessments
Review group assessments
Summary of Round 1 Estimates
Aggregated
0%
20
60
80
100%
Best estimate of the probability of successful replication (%)
Lower bound (%)
Upper bound (%)
20
60
80

Figure 1. The IDEA protocol as supported by the repliCATS platform

permanently saved when they are uploaded via the bulk upload process while the latter only have their accounts saved after they complete the consent and demographics steps described below.

All participants are automatically assigned avatar images and avatar names on participant creation in order to anonymise judgements and comments as far as possible. This is a feature of the IDEA protocol, as described above.

### 5.2. Consent and demographic data collection

The repliCATS platform collects consent and basic demographic data from all participants prior to first use by integrating with a Qualtrics questionnaire. Some answers to the Qualtrics questionnaire are mandatory (eg consent) while other non-essential demographics fields are optional. This information is used along with the claim assessments in aggregating participant judgements and informing the qualitative and quantitative analysis.

## 6. Platform claim management

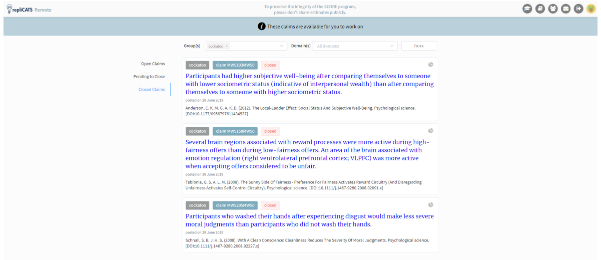
The core of the repliCATS project is to produce assessments of individual research claims. Within the SCORE program, a research claim is a single finding from within a published paper that is based on a specific quantitative test. Such claims are quite granular and typically refer to a single entry within one table of the paper. An example claim is that ”Participants answered more moderately difficult syllogisms correctly when the font was hard to read than when it was easy to read”, specifically tested through difference between accuracy rates for the two conditions. Research claims come from a wide variety of disciplines and include both experimental results and statistical modelling and analysis of existing datasets. Each such claim is described by: text describing the claim and the inferential test used for it; statistical parameters for the inferential test, including sample size, effect size and p-value (where available); as well as the original paper. These individual research claims and their metadata are drawn from a library of research claims, which is not developed by the repliCATS project but supplied to it by the TA1 of the SCORE program. Thus, the repliCATS platform is required to be able to both import and display this externally-supplied metadata and link to an external repository of papers for scrutiny by participants. The repliCATS platform collects the quantitative judgement data for aggregation into group assessments as well as all comments from participants, considered to be reasoning data for these judgements.

## 6.1. Platform management of claim pool

Discussing the SCORE TA1's method for creating this claims pool is beyond the scope of this paper. The output of this activity, however, is a JSON file that includes research papers, specific claims identified within those papers and associated metadata for those research claims, such as statistical details. Claims from the claim pool are imported into the repliCATS platform through an API designed to consume the TA1 JSON file format. The API is triggered manually by a participant with administrative access via a Python script. An additional upload file allocates claims to groups within the repliCATS platform. An individual claim can only be assigned to a single group (although it is possible to make a duplicate of a claim and assign it to multiple groups). For example, all participants operating in Remote mode are assigned to a single group, and all claims to be assessed remotely are allocated to this group.

## 6.2. Claim selection across both modes

The main landing page for participants who have been enrolled and completed consent and demographics is the claim selection page. The key features of the claim selection page include: (1) claim summary details and hyperlink; (2) claim progress counter; (3) claim sorting; and (4) claim queues. There is no dedicated claim search function; the platform replies on the web browser's search functionality. A screenshot of the repliCATS platform claim selection page is shown in Figure 2.



To preserve the integrity of the SCORE program, please don't share estimates publicly.

These claims are available for you to work on

cockatoo | claim MWS103MWS6 | closed

Participants had higher subjective well-being after comparing themselves to someone with lower sociometric status (indicative of interpersonal wealth) than after comparing themselves to someone with higher sociometric status.

posted on 28 June 2019

Anderson, C. K. M. G. A. K. D. (2012). The Local-Ladder Effect: Social Status And Subjective Well-Being. Psychological science. [DOI:10.1177/0956797611434537]

cockatoo | claim MWS158MWS6 | closed

Several brain regions associated with reward processes were more active during high-fairness offers than during low-fairness offers. An area of the brain associated with emotion regulation (right ventrolateral prefrontal cortex; VLPFC) was more active when accepting offers considered to be unfair.

posted on 28 June 2019

Tabibnia, G. S. A. L. M. (2008). The Sunny Side Of Fairness - Preference For Fairness Activates Reward Circuitry (And Disregarding Unfairness Activates Self-Control Circuitry). Psychological science. [DOI:10.1111/j.1467-9280.2008.02091.x]

cockatoo | claim MWS205MWS6 | closed

Participants who washed their hands after experiencing disgust would make less severe moral judgments than participants who did not wash their hands.

posted on 28 June 2019

Schnall, S. B. J. H. S. (2008). With A Clean Conscience: Cleanliness Reduces The Severity Of Moral Judgments. Psychological science. [DOI:10.1111/j.1467-9280.2008.02227.x]

**Figure 2. Example of the claim selection page of the repliCATS platform**

### 6.2.1. Claim summary details and hyperlink

All claims display information, including: claim id, a hyperlink title, the group the claim is assigned to, related disciplines, the date added to the platform and the associated article reference. Clicking the hyperlink title takes the participant to the assessment page.

### 6.2.2. Claim progress counter

An icon in the top right corner, upon hovering, provides a count and list of participants that have: (1) access to the claim; (2) commenced first round; (3) completed the first round but not the second round; and (4) completed the second round. This information is intended to help participants in selecting claims that are relevant to them while informing them of other participants' activity. An example of a progress counter is shown in Figure 3.

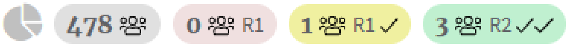


**Figure 3. Example of a claim progress counter showing: 478 participants with access to the claim: 0 participants who have started (but not completed) a first round judgement; 1 participant who has completed a first round (but not a second round) judgement; and 3 participants who have completed a second round judgement**

### 6.2.3. Claim sorting

The claim selection page sorts the claims using primary and secondary sorting heuristics. The primary heuristic sorts the claims by round, to help participants track the progress of claims they have completed. Claims in first round appear first followed by second round claims. The primary heuristic is most useful for workshops. The secondary heuristic sorts by the number of assessments made by other participants per claim in descending order so that claims with the most responses appear at the top. This encourages participants to finalise claim assessments more quickly, and ensures they see the activity of other participants. This heuristic is primarily appropriate for Remote mode.

### 6.2.4. Claim queues

Claims exist within one of three separate queues: 'Open Claims', 'Pending to Close' and 'Closed Claims'. These relate to how claims are closed in the Remote mode workflow and will be discussed in the Claims Closing section below.

## 6.3. Assessing claims with the IDEA protocol

Claim assessment on the repliCATS platform is a two round process which implements the IDEA protocol. Both rounds allow the participant to review the claim and answer specific questions about the replicability of that claim. The assessment screen in the first round displays two panels: the *claim details* *panel* and the *claim questions panel*. The second round screen allows the participant to view other participant's judgements and justifications, and to revise their own. The assessment screen in the second round round displays three panels with a *responses and justifications panel* shown alongside the claim details and the claim questions panels.

The *claim details* panel provides the participant with information relating to the claim, such as the statistical test that is actually being evaluated for replicability, as derived from the metadata supplied by TA1, and a link to the actual paper which is accessible in an external online repository hosted on the Open Science Foundation (OSF) website. These details enable participants to Investigate the claim. The claim details panel is persistent and unchanged between the first and second rounds. The definition of the elements found in the claim details panel can be found in Appendix B[2].

The *claim questions panel* lists a series of questions. The question wording and sequence is designed to elicit both quantitative and qualitative judgements about the replicability of the claim while minimising the impact of potentially confounding aspects, like the clarity or usefulness of the paper. A full description of the design of these questions is beyond the scope of this paper, but the specifications of the questions presented in both rounds can be found in Appendix C[3]. Response types for these questions are: Likert scale; binary (i.e. yes/no), three-step quantitative elicitation (i.e. lower bound, upper bound, best estimate) and free-text. Some questions have two response types, including a free-text box, while others have only one. Free-text boxes catch the rich qualitative justifications that are a key feature of the repliCATS project approach.

Once all mandatory responses are entered, the judgement can be submitted by pressing the submit button on the claim question panel. (Free-text responses are never compulsory; all quantitative responses are.) The submission button invokes two different participant experiences depending on workflow mode. In Workshop mode, 'submit' returns the participant to the claim selection page and moves the submitted claim from the first round to the second round. This reflects the experience that workshop groups often complete multiple first round assessments in a row before returning to complete second round assessments with in-group discussion and returning to the claim selection panel retains flexibility for that behaviour. In Remote mode, 'submit' also moves the claim from the first round to the second round. However, the participant remains in the claim assessment screen and is prompted to immediately complete their second round assessment. This encourages participants to complete both rounds in a single session, for reasons described below.

The format of the claim questions panel is similar between the first and second rounds. (One question asked in the first round is a factual one that requires no updating.) The question wording for the second round questions is slightly different to reflect that they ask about revised judgements rather than original ones, but the basis of those questions is unchanged.

The differences in the claim questions panel between the two rounds allows for the implementation of the Discussion phase of the IDEA protocol. One such difference is that the judgements of other participants for all questions is displayed in in the second round in the claim questions panel, along with a simple group aggregate (i.e. arithmetic mean) of the quantitative assessment of the replicability of the research claim. The other main difference is the *responses and justifications panel* that appears in the second round, containing the free-text responses from all participants for the first round. This new panel also enables dialogue between participants to be captured in the form of comments, threads and up-voting. (In Workshop mode, there is typically a face-to-face discussion phase. Describing how this discussion is captured and analysed is beyond the scope of this paper, but even in this mode, participants are encouraged to record important aspects of this discussion in comments and second round justifications in order to provide the rich qualitative data.) Both differences allow participants to consider information, reasoning and judgements provided by other members of the group. In turn, this encourage participants to reflect on their own responses. The IDEA protocol does not require participants to change their judgements for the sake of change. However, it does encourage them to re-think judgements based on potentially new information.

## 6.4. Claim closure to allow aggregation

Claims need to be 'closed' at some point in time so that assessments of the claim can be finalised. The way this is done differs in the two workflow modes. In Workshop mode, there is no pre-set time limitations imposed on claims. All claims are left open for participants to revise their judgements throughout the workshop period (which is typically one or two days). After the workshop is completed, claims are closed manually by repliCATS platform administrators.

This process is more complex in Remote mode. Because claims are self-selected and assessed asynchronously, there is no way of knowing when any

[2] Appendix B - Claim Details Panel - https://osf.io/qu27k/

[3] Appendix C - Round Question Wording - https://osf.io/qu27k/

given claim will have attracted sufficient judgements for a finalised assessment. Thus, the platform needs to automatically close the claim after it has attracted a sufficient number of judgements. A drawback of this is that the first participant to complete the first round will not be able to consider any other participants' judgements before submitting their second round judgements, and the second participant will only have the benefit of one judgement, and so forth.

The 'Pending to Close' queue was developed for the Remote mode workflow to mitigate this. Rather than claims closing immediately after the final judgement is entered, there is a 72-hour window in which the claim is closed to new participants, but open to existing participants to review the judgements of other participants and revise their own judgements if desired. Furthermore, participants are encouraged to do so by a transactional e-mail sent indicating that the claim is 'Pending to Close'.

Once the 72 hour window expires, the claim is moved to the 'Closed Claim' queue. In this state it is set to read-only for participants, who can revisit their past assessments of claims but cannot change them. The finalised group assessment can then be aggregated.

Determining the optimal number of judgements before a claim is put into the 'Pending to Close' queue is not straightforward. The IDEA protocol is based on the understanding that judgements are improved through the provision of additional information and reasoning, but it is not clear exactly how much additional information is useful. The repliCATS project initially used a group size of five for Workshop mode, based on our interpretation of the existing evidence base [23]. After 150 claims had been assessed in Remote mode, the decision was taken to reduce the number of individual judgements to four. While this decision was taken on purely pragmatic grounds, we were interested to determine the effect this change had on the quality of assessments. Using the change in judgements between the first and second rounds as a proxy for quality assessment, this investigation suggested that a group size of four was feasible, but that groups below four were not desirable. A discussion of this investigation is provided in Appendix D[4].

## 6.5. Claims aggregated and reported

All participant assessments, justifications and discussions are recorded in the repliCATS platform database. The repliCATS platform provides daily export files that include participants' quantitative and qualitative judgements about claims in a usable JSON format. These export files are then imported into downstream analytics for aggregation.

# 7. Platform engagement of participants

The participant base for the platform is large and diverse. Undertaking assessments of 3000 claims with a minimum group size of four individual judgements per claim requires a substantial participant community with consequent recruitment and engagement challenges.

## 7.1. Users notified of events

One challenge is keeping the large repliCATS platform participant base notified of significant events. This is done via email. There are two forms of emails, bulk and transactional. Examples of these, as above, include emails to newly enrolled users – bulk emails for bulk upload users and transactional emails for self-enrolled users – and, in Remote mode, transactional emails to relevant participants when a claim is moved to 'Pending to Close'. The repliCATS platform integrates with the Mailchimp and Mandrill mail platforms via API; it does not directly support Simple Mail Transfer Protocol (SMTP) or other mail protocols.

## 7.2. Decision support materials provided

The academic background of participants varies considerably and not all have the same level of knowledge and expertise in replication studies. Moreover, the assessments involve decisions about how to measure replicability, e.g. what counts as a successful replication, that are non-obvious and therefore will not be shared amongst the participant community prior to engaging with the platform. As a result, there is a need for decision support materials provided in a user-friendly manner that can provide clear information about these matters.

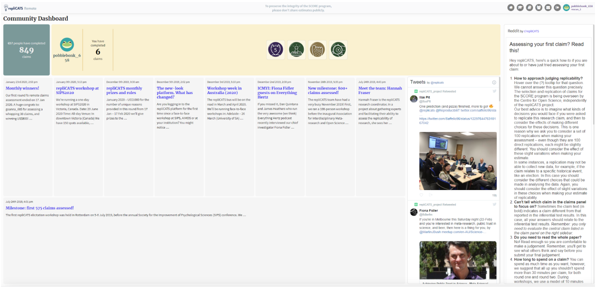
There are several sources of material that support repliCATS platform participants. During enrolment, the Qualtrics consent and demographic questionnaire provides training. In-platform, there are tooltips for each question to provide specific detail about the intent of the question and how technical terms are defined and used in the repliCATS project, and free-text boxes contain placeholder text with suggestions and examples of possible responses, prior to text entry. Hosted on the repliCATS project website – and directly linked from the platform – is a wide variety of resources including a glossary of technical terms, user guides for the repliCATS platform, and videos and short e-courses on replication and statistical concepts frequently used in the research claims under assessment.

[4] Appendix D - Group Size Assessment - https://osf.io/qu27k/

### 7.3. Community engagement and gamification

The recruitment and engagement challenges for the repliCATS project are also substantial. Both intrinsic and extrinsic motivations of participants in academic research projects are highly variable and attrition rates from fatigue need to be considered. Additionally, Remote mode has significant challenges from a participant engagement perspective, as it relies upon participants' self-motivation.

A Community page was developed to support community engagement, particularly in the context of Remote mode. This page is an externally hosted website that is directly linked from the repliCATS platform. Elements of this page include a universal claim counter, an individual participant claim counter, live news and posts, an embedded twitter feed, a Reddit support page, and finally participation-based badges. The purpose of this page is to help participants navigate or re-orient themselves with the platform and the task of evaluating a claim; create and retain a sense of community in a fully virtual environment; and house gamification elements for rewarding participation, such as participant badges. The last of these emphasises behaviours considered valuable to the elicitation process, for example, regularly submitting verbal reasoning, accessing decision support materials, and engaging with other participants' judgements by upvoting and comments. Figure 4 shows a screenshot of the repliCATS platform Community page.



To preserve the integrity of the SCORE program, please don't share estimates publicly.

Community Dashboard

657 people have completed 849 claims

pobblebonk_658

You have completed 6 claims

January 23rd 2020, 2:53 pm
**Monthly winners!**
Our first round fo remote claims assessment ended on 17 Jan 2020. A huge congrats to: goanna_685 for assessing a whopping 30 claims, and winning US$800! ...

January 9th 2020, 5:13 pm
**repliCATS workshop at SIPS2020**
We're running a one-day workshop at SIPS2020 in Victoria, Canada. Date: 20 June 2020 Time: All-day Venue: In downtown Victoria (Canada) We have 150 spots available, ...

December 9th 2019, 9:30 am
**repliCATS monthly prizes and rules**
January 2020 – US$1000 for the number of unique reasons provided in this round from 17 Jan – 17 Feb 2020 we'll give prizes to the ...

December 5th 2019, 2:02 pm
**The new-look platform. What has changed?**
Are you logging in to the repliCATS platform for the first time since a face-to-face workshop at SIPS, AIMOS or at your institution? You might notice ...

December 3rd 2019, 5:10 pm
**Workshop week in Australia (2020)**
The repliCATS bus will be on the road in March and April 2020. We'll be running face-to-face workshops in: Adelaide – 26 March (University of SA) ...

December 2nd 2019, 2:38 pm
**ICMYI: Fiona Fidler guests on Everything Hertz**
If you missed it, Dan Quintana and James Heathers who run the very awesome (we think) Everything Hertz podcast recently interviewed our chief investigator Fiona Fidler ...

November 26th 2019, 5:35 pm
**New milestone: 800+ claims assessed!**
The repliCATS team have had a very busy November 2019! First, we ran a 100-person workshop before the inaugural Association for Interdisciplinary Meta-research and Open Science ...

July 24th 2019, 4:43 pm
**Meet the team: Hannah Fraser**
Hannah Fraser is the repliCATS research coordinator. In a project about gathering experts and facilitating their ability to assess the replicability of research, she sees her ...

July 24th 2019, 4:01 pm
**Milestone: first 575 claims assessed!**
The first repliCATS elicitation workshop was held in Rotterdam on 5-6 July 2019, before the annual Society for the Improvement of Psychological Sciences (SIPS) conference. We ...

Tweets by @replicats

repliCATS_project Retweeted
Ilse Pit
@IlsePit
One prediction (and pizza) finished, more to go! 🔥 @replicats @ReproducibiliT twitter.com/saffello96/sta ...
https://twitter.com/Saffello96/status/1229764476318167042

19h

repliCATS_project Retweeted
Fiona Fidler
@fidlerfm
If you're in Melbourne this Saturday night (22-Feb) and you're interested in meta-research, public trust in science, and beer, then here is a thing for you, by @MartinJBush meetup.com/en-AU/Science-...

Achieving Public Trust in Science - Meta-Science!

Reddit by r/repliCATS

Assessing your first claim? Read this!

Hey repliCATS, here's a quick how-to if you are about to or have just tried assessing your first claim:

1. **How to approach judging replicability?** Hover over the (?) tooltip for that question. We cannot answer this question precisely. The selection and replication of claims for the SCORE program is being overseen by the Centre for Open Science, independently of the repliCATS project.
   Our best advice is to imagine what kinds of decisions you would face if you were asked to replicate this research claim, and then to consider the effects of making different choices for these decisions. This is one reason why we ask you to consider a set of 100 replications when making your assessment – even though they are 100 direct replications, each might be slightly different. You should consider the effect of these slight variations when making your estimate.
   In some instances, a replication may not be able to collect new data, for example, if the claim relates to a specific historical event, like an election. In this case you should consider the different choices that could be made in analysing the data. Again, you should consider the effect of slight variations in these choices when making your estimate of replicability.
2. **Can't tell which claim in the claims panel to focus on?** Sometimes the claim text (in bold) indicates a claim different from that reported in the inferential test results. In this case, all your answers should relate to the inferential test results. Remember: *you only need to evaluate the central claim listed in the claim panel on the right sidebar.*
3. **Do you need to read the whole paper?** No! Read enough so you are comfortable to make a judgement. Remember, you'll get to see what others think and say before you submit your final judgement.
4. **How long to spend on a claim?** You can spend as much time as you want, however, we suggest that all up you shouldn't spend more than 30 minutes per claim, for both round one and round two. During workshops, we use a model of 10 minutes

**Figure 4. The repliCATS platform Community page**

## 8. repliCATS platform architecture

The architecture for the repliCATS platform was modelled on the SWARM platform [24]. This was primarily for pragmatic reasons as the repliCATS project utilised the same development team, development pipeline and code base of SWARM. The decision to branch from an existing SWARM code base rather than develop a new one was motivated by the aggressive timelines of the program and the need to start collecting claim assessments as early as possible. The repliCATS platform and SWARM shared significant commonality in purpose which made the branching feasible. Despite the origin of the code base, the repliCATS platform end product deviates markedly from the SWARM platform. Of final note is that in addition to the code base, some User Interface elements have also been retained such as avatar names and avatar logos.

The repliCATS platform architecture adopts a microservice centric approach with the definition of a microservice (for the purpose of this paper) being small independently deployable services that are decentralized and adopt a DevOps CI/CD (CI / CD Continuous Integration and Continuous Deployment) approach to implementation. Figure 5. shows the overall architecture for the repliCATS platform including each element of the solution as well as their integration.

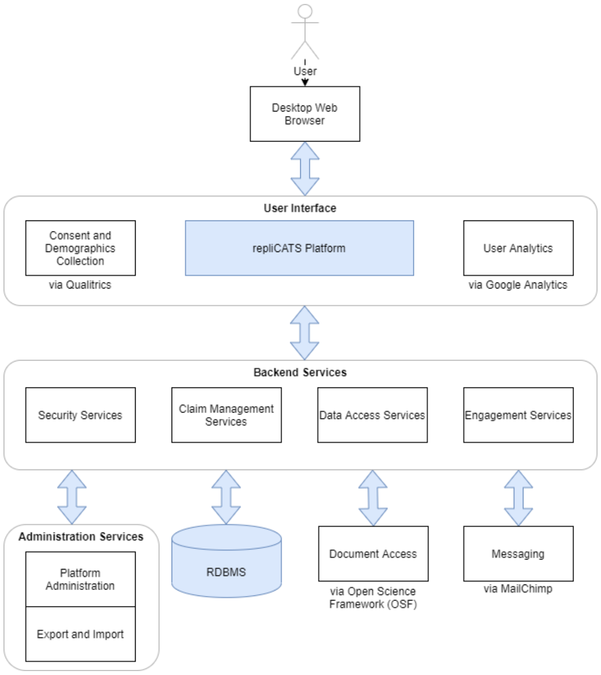


**Figure 5. repliCATS Platform System Architecture**

The repliCATS platform is designed using contemporary software engineering approaches and delivered as a rich internet application (RIA). By intent, wherever possible, user interactions are executed on the client side through a web browser. Elements of data validation and business logic are performed within the UI components of the solution to distribute load between the client and server. The repliCATS platform is delivered to the client as a single page application (SPA), where data and core business logic are retrieved/executed from the server on-demand and only when needed by the user.

The core microservice components serve logically related-services, including: (1) security services; (2) data access services; (3) claim management services; and (4) engagement services. Benefitting from cloud-based Software as a Service (SaaS), the repliCATS platform also leverages external services including: (1) Qualtrics, (2) Mailchimp (and Mandrill), and (3) Google Analytics, to speed up the delivery and increase reliability through third-party provisioning.

The core backend is developed as a ReST-based API provider. This API layer provides services to various components (security, data access, claim and engagement) as well as providing a data abstraction layer between user and physical storage in the Relational Database Management System (RDBMS).

The data access component provides a means to access the collection of claims, supporting claim metadata and both quantitative and qualitative responses. This data access is wrapped by the security component, which provides authentication and authorization services for data based on JSON Web Tokens (JWT) that provide a flexible mechanism to access APIs in the web context. An administration module allows the platform manager to manage user registration and allocation, account distribution and system configuration settings. Batch claim data ingestion can also be done through this module. This supports the export of both operational and user interaction data for more advanced data analytics tools.

The main purpose of the backend is to facilitate the IDEA protocol through state management of the claims and the relevant user responses related to the claims being assessed. These roles are played by the claim management component. This component supports the complete execution of the IDEA protocol workflow starting from claim allocation, individual contribution, collective contribution and online discussion before final progression towards a replicability consensus.

At the same time, data analytics are carried out and business rules applied to drive claim completion in general and to invoke participant engagement. In addition, the engagement component collects usage statistics to spur gamification elements such as participant badge determination and/or recognition mail notifications delivered through mailchimp.

## 9. Future work and conclusion

The repliCATS platform has demonstrated its usability in multiple settings. At time of writing, over a period of 15 months, 637 individual participants have used the platform to make more than 17000 judgements on 3432 claims. The platform has shown the required scalability for the research question described above.

Accuracy data for assessments of SCORE claims are not yet available. The work described in this paper goes to the viability of the approach. The extent of its fruitfulness is for the future but clearly comparisons of the performance of this technique to other approaches is of interest.

Once accuracy data is available, a number of research questions will be opened to the repliCATS project. The extent of these is described more fully in other papers forthcoming from the repliCATS project. They include include interrogating the unique qualitative dataset to see if styles of reasoning are associated with improved accuracy and analysing demographics data for associations that can be tested.

Potential applications of the current platform include developing collaborative assessment as a form of alternative peer-review of papers or projects, and allocation of replication resource effort. With technical improvements to the repliCATS platform, even more problems will become tractable. As previously, the field of replicability research is accelerating and there are further questions pertinent to reliability of research claims, including their generalizability and translatability. Addressing these is currently limited by a lack of a generic API for importing claims or modifying question types; the current platform was designed for a specific elicitation. An API with content management overlay supporting the import of claims from multiple sources and a library of question response types (e.g. Likert scale, three-point quantitative, etc) will substantially extend the research potential of the platform. The usability of the repliCATS platform, too, can be enhanced based on feedback from users.

The implementation of the IDEA protocol on the repliCATS platform is a novel approach with potential to make substantial gains for replicability research within the IS field and beyond. A feasible implementation of a computer-mediated human judgement requires careful consideration of theoretical, technical and user-oriented features. Current results suggest the repliCATS platform is able to both scalably generate confidence scores in published literature while contributing to seveal other significant research problems.

## 10. Acknowledgement

This research is based upon work funded by the Defence Advanced Research Projects Activity (DARPA), under Contract HR001118S0047 as part of the SCORE program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

## References

[1] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, pp. aac4716–aac4716, Aug. 2015.

[2] C. G. Begley and L. M. Ellis, "Raise standards for preclinical cancer research," *Nature*, vol. 483, pp. 531–533, Mar. 2012.

[3] C. F. Camerer, A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu, "Evaluating replicability of laboratory experiments in economics," p. 5, 2016.

[4] A. C. Chang and P. Li, "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not"," *Finance and Economics Discussion Series*, vol. 2015, pp. 1–26, Oct. 2015.

[5] F. Cova, B. Strickland, A. G. F. Abatista, and Z. Xiang, "Estimating the Reproducibility of Experimental Philosophy," *Review of Philosophy and Psychology*, vol. In press, June 2018.

[6] C. F. Camerer, A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell, A. Gampa, E. Heikensten, L. Hummer, T. Imai, S. Isaksson, D. Manfredi, J. Rose, E.-J. Wagenmakers, and H. Wu, "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015," *Nature Human Behaviour*, vol. 2, pp. 637–644, Sept. 2018.

[7] J. K. Hartshorne and A. Schachner, "Tracking replicability as a method of post-publication open evaluation," *Frontiers in Computational Neuroscience*, vol. 6, p. 8, 2012.

[8] N. A. Vasilevsky, M. H. Brush, H. Paddock, L. Ponting, S. J. Tripathy, G. M. Larocca, and M. A. Haendel, "On the reproducibility of science: unique identification of research resources in the biomedical literature," *PeerJ*, vol. 1, p. e148, 2013.

[9] M. C. Makel, J. A. Plucker, and B. Hegarty, "Replications in Psychology Research: How Often Do They Really Occur?," *Perspectives on Psychological Science*, vol. 7, pp. 537–542, Nov. 2012.

[10] C. D. Kelly, "Rate and success of study replication in ecology and evolution," *PeerJ*, vol. 7, p. e7654, Sept. 2019.

[11] A. Dennis and J. Valacich, "A Replication Manifesto," *AIS Transactions on Replication Research*, vol. 1, pp. 1–4, Aug. 2014.

[12] A. Cockburn, P. Dragicevic, L. Besançon, and C. Gutwin, "Threats of a replication crisis in empirical computer science," *Communications of the ACM*, vol. 63, pp. 70–79, July 2020.

[13] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel, "Improvements that don't add up: ad-hoc retrieval results since 1998," in *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*, (Hong Kong, China), p. 601, ACM Press, 2009.

[14] F. Niederman and S. March, "Reflections on Replications," *AIS Transactions on Replication Research*, vol. 1, pp. 1–16, 2015.

[15] J. Thatcher, W. Pu, and D. Pienta, "IS Information Systems a (Social) Science?," *Communications of the Association for Information Systems*, pp. 189–196, 2018.

[16] R. K. Stamper, "Information Systems as a Social Science," in *Information System Concepts: An Integrated Discipline Emerging* (E. D. Falkenberg, K. Lyytinen, and A. A. Verrijn-Stuart, eds.), vol. 36, pp. 1–51, Boston, MA: Springer US, 2000. Series Title: IFIP Advances in Information and Communication Technology.

[17] A. Dennis, S. Brown, T. Wells, and A. Rai, "Replication Crisis or Replication Reassurance: Results of the IS Replication Project," *MIS Quarterly*, vol. 44, pp. iii–vii, Sept. 2020.

[18] A. Dreber, T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, and M. Johannesson, "Using prediction markets to estimate the reproducibility of scientific research," *Proceedings of the National Academy of Sciences*, vol. 112, pp. 15343–15347, Dec. 2015.

[19] Y. Yang, W. Youyou, and B. Uzzi, "Estimating the deep replicability of scientific findings using human and artificial intelligence," *Proceedings of the National Academy of Sciences*, p. 201909046, Apr. 2020.

[20] A. M. Hanea, M. F. McBride, M. A. Burgman, and B. C. Wintle, "The Value of Performance Weights and Discussion in Aggregated Expert Judgments: The Value of Performance Weights and Discussion," *Risk Analysis*, vol. 38, pp. 1781–1794, Sept. 2018.

[21] V. Hemming, M. A. Burgman, A. M. Hanea, M. F. McBride, and B. C. Wintle, "A practical guide to structured expert elicitation using the IDEA protocol," *Methods in Ecology and Evolution*, vol. 9, pp. 169–180, Jan. 2018.

[22] A. Hanea, M. McBride, M. Burgman, B. Wintle, F. Fidler, L. Flander, C. Twardy, B. Manning, and S. Mascaro, "I nvestigate D iscuss E stimate A ggregate for structured expert judgement," *International Journal of Forecasting*, vol. 33, pp. 267–279, Jan. 2017.

[23] F. Fidler, B. Wintle, and N. Thomason, "Groups Making Wise Judgements," tech. rep., National Office of the Director of National Intelligence, 2013.

[24] R. Sinnott, "The Design and Development of a Cloud-based Platform Supporting Team-oriented Evidence-based Reasoning: SWARM Systems Paper," 2019.