Guide to making assessments on the bushel platform

Introduction

This document contains a quick reminder of the aims and approach of the repliCATS project, followed by a guide to answering the claims in Round 1 and Round 2 of the IDEA protocol. There is also a list of FAQs.

The repliCATS project

The University of Melbourne repliCATS team elicits expert judgements about the credibility of research claims in the Social and Behavioural Sciences through an online platform using the IDEA protocol. Judgements are aggregated into measures of reliability and the reasoning used is analysed. IDEA ("Investigate", "Discuss", "Estimate" and "Aggregate") has been found to improve judgements under uncertainty. More information about the repliCATS project is contained in the Plain Language Statement and on the repliCATS website.

The IDEA protocol

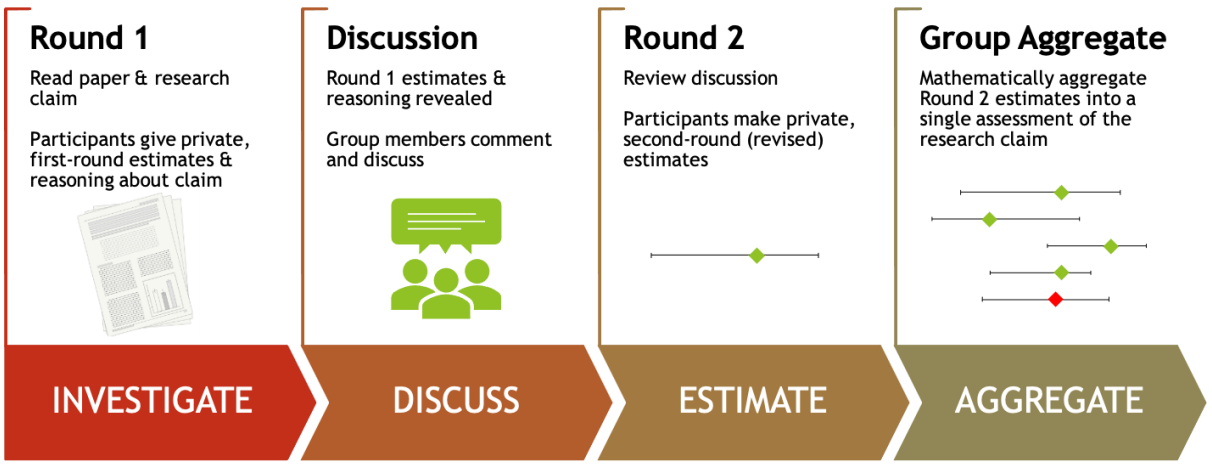
This protocol, developed at the University of Melbourne, has been found to improve judgements under uncertainty. IDEA stands for "Investigate", "Discuss", "Estimate" and "Aggregate", the four steps in the process of this elicitation.

As used in the repliCATS project, the IDEA protocol will involve participants:

1. Independently *Investigating* the claim, providing their personal judgement on the credibility of the claim, and commenting on their thinking.
2. Seeing the judgements of the rest of their team, the aggregated judgement, and all of the comments that have been made, and having a *Discussion* with the group. This phase can resolve uncertainties and investigate evidence and thinking.
3. Providing a revised *Estimate* and describing how their thinking has changed.

The repliCATS team will use an *Aggregate* of the group judgements as the final assessment of the replicability of the research claim.

More information on the IDEA protocol can be found here.



Guidelines for Round 1

Entering your Round 1 estimates

a) **Login** to the repliCATS platform at https://replicats-bushel.eresearch.unimelb.edu.au. If you have difficulty, please contact replicats-contact@unimelb.edu.au. The platform is tested for Google Chrome on laptops. It may not work in incognito mode, on other browsers, or on mobile devices.

b) **Select a claim.** Upon login you will see the list of claims that have been assigned to your group. Select any claim by clicking on it. You'll be taken to the claim assessment page with information about the claim and the elicitation questions (outlined below).

c) **Read the paper.** The platform provides a link to the original paper. As we are assessing the credibility of the paper, we expect you to have read a substantial amount of the paper. Some papers are very long and detailed. It may not be necessary to go through all information exhaustively. However, you need to read enough to assess all aspects of the paper fairly. We don't expect you to take more than about 30 minutes on this task.
We encourage you to use whatever information you would like to inform your judgements. You can speak to others and search for additional sources of evidence. However, if you know someone is assessing the same paper as you, please ***do not*** discuss your judgements with them – you can do this in the *Discussion* phase.
Please note that the papers have been selected by a team that is independent to repliCATS. A small number of published papers contain inappropriate or potentially distressing material. We aim to remove such papers. If you think that you have been assigned such a paper, please stop reading it and notify a repliCATS team member.
In the unlikely event that you have been assigned a paper which you are involved in attempting to replicate, please do not enter any assessments, and notify the workshop co-ordinator or another repliCATS team member.

d) **Answer the paper-level questions.** Details of these questions are provided below. We don't expect you to spend longer than 10-15 minutes per claim answering these questions. You will have time to revise your estimates in Round 2.
We encourage you to document all of your reasoning in the comment box for *Question 8: Credibility*. There are two other comment boxes - pl*ease use these only for the specific purpose suggested.* We are also interested in understanding how you thought about the paper and what characteristics were important for your judgement with regards to any of the questions. Note that all comments will be attributed to an avatar. ***Please do not use your own name or the name of any other participants.*** Judgements will eventually be made public and we must be able to keep them anonymous.

e) **Answer the evidence-level questions.** After you have answered the paper-level questions, you will be directed to more detailed questions about specific claims that relate to individual test results within the paper. There will be a variable number of these questions per paper - some papers may have only a few, some may have up to ten. We expect that this should take about 5 minutes to complete. Details of these questions are provided below. Again, you will have time to revise your estimates in Round 2.
If you wish to make any comments about any of these evidence-level questions, please go back to the comments box for *Question 8: Credibility* and add them there. Note that you can navigate to this comments box and update it at any time.

f) **Submit your estimates**. 'Saving' your estimates will enable you to return to the estimates

and update them prior to submission. Pressing ‘submit’ finalises your estimates for Round 1 - but you’ll have an opportunity to revise these in Round 2 following the Discussion phase. You can only submit your Round 1 judgements once you have answered all of the questions. Once your Round1 judgements have been submitted, the paper should be listed as ‘Round 2’ on the home page. If not, you will need to click back into the paper, check that all answers are filled in, and re-submit.

That’s Round 1 completed. You can also see the basic steps in completing Round 1 on the platform in a video that can be found here.

Answering the questions

Assessing a paper on this version of the repliCATS platform involves considering both features of the overall paper, as well as some specific claims within the paper. Questions 1-8 refer to paper-level assessments. Following that, you will be asked to consider a variable number of specific results per paper (no more than 10 per paper), answering additional questions per specific result.

Paper level questions

In this version of the repliCATS platform, we are trying to assess the credibility of scientific papers as a whole. There are many dimensions to credibility and so we need to ask you questions about different aspects of the paper. Some questions and some credibility dimensions may seem linked and you might find distinguishing between them tricky. The tool tips on the platform and this guide should help you differentiate between the credibility dimensions and answer the questions as accurately as possible.

When you answer the paper-level questions, we want you to think about what you understand the main claim of the paper to be. By that we mean the key finding, which typically would be described in the paper’s abstract and highlighted in the paper’s conclusion. However, if you just want to think about your overall impression of the paper, that is fine. We are aware that different people might form different opinions about what the central claim of a given paper is. That’s okay. You can discuss this with your group after you have completed Round 1.

Question 1: Comprehensibility

How well do you understand the paper overall?

Purpose: To understand if anything affects your ability to interpret the paper and identify its central claim.

Clarification: We’ve done our best to make sure the papers are interpretable but we know they vary in clarity and comprehensibility. It’s possible that the paper:

- is vague;
- is poorly written;
- relies on an unfamiliar procedure;
- contains too much jargon;
- is unclear about exactly what the *central* claim is;
- is about a concept that you are not familiar with and/or have difficulty conceptualising.

These factors can all contribute to your ability to be able to interpret the claim and may in turn lead to different interpretations by the group. There is a comments box below this question,

where you can provide a summary of the paper's central claim, as you see it. We would like you to focus on the higher-level finding (the take-home message, if you will), not detailed results. There will be space to consider individual results later. Sharing your interpretation will help highlight whether there are different opinions about the central claim of the paper, and this information may be useful during the Discussion.

Answering the comprehensibility question:

We're asking for this on a scale of 0 to 100. 0 means that you have no idea what the paper means, 100 means it's perfectly clear to you. This is not an objective measurement - try your best to estimate a number for comprehensibility, and try to be consistent between papers.

Some papers may be outside of your main fields or use words that you are unfamiliar with. This might cause you to immediately put 'I have no idea what the paper means'. However, with a little bit of effort you can usually deduce what is being asked. If after completing your reading of the paper you still cannot work it out then you should definitely indicate this to us and consider whether that is indicating something about the quality of the research being described.

Tooltip: We know the clarity and comprehensibility of scientific papers varies. We are interested in your honest account of how well you understand this paper and its central claim

Comments box:

We have provided a comment box below the question so you can try to rephrase what you think the paper is about and what the central claim is. This will be really useful in the Discussion phase to prompt your memory about your initial interpretation of the question. *Please do not place any discussion about your assessment off the paper here* - put all such comments in the comments box for *Question 8: Credibility*. Note that you can navigate to this question at any stage to add comments as you go - you don't need to wait until you start assessing Question 8.

The placeholder text for this question is:

> *Please restate the central claim of this paper and highlight terms, concepts or other features of the paper that make it hard to understand.*

Question 2: Plausibility

How consistent is the central claim of this paper with your existing belief?

Purpose: To capture your beliefs about whether the underlying effect or relationship corresponds to something real.

Clarification: Sometimes we hear a series of claims in a paper and we have a strong feeling that some or all of them do not seem very plausible either within the context of the experimental design, or more broadly (i.e. relating to a relationship that would generalise across contexts, or experimental designs).

These prior beliefs can be useful. We've included this question here to allow you to state your prior belief about if you think there is a real effect in this study, regardless of what you think about this particular experimental/study design.

Don't spend too much time on this question. In the next question, we want you to examine the claim and the validity of your prior beliefs more critically, as to how they relate to direct

replication.

Answering the plausibility question:

We're asking for this on a scale of 0 to 100. 0 means that the paper is exactly contrary to your pre-existing beliefs, 100 means it's perfectly compatible with them. As with *Question 1: Comprehensibility*, try to estimate a number and try to be consistent between papers.

The word 'plausible' means different things to different people. For some people almost everything is 'plausible', while other people have a stricter interpretation. Don't be too focused on the precise meaning of 'plausible' – you could also consider words like 'possible' or 'realistic' here. We just ask you to maintain a consistent standard between different claims and try to let us know if some papers are clearly *more* plausible (or implausible) than others.

If you didn't understand the paper being asked, it might be challenging to say whether you believe it's plausible. Hopefully, the paper will become clearer in the Discussion phase.

Tooltip: We know this is a broad question. Consider how you would have assessed the plausibility of the central claim before you read the paper. We're trying to get a sense of your degree of belief in whether the effect or relationship underlying the claim is real.
You will have an opportunity to comment on your reasons for this in question 8. (Credibility) below.

Question 3: Transparency

Based on your quick read of the paper, how transparent is the research described here? Think about how easy or difficult it would be for someone who wanted to evaluate or replicate the research in this paper to find all the information they need about the methods, analysis and procedures.

Purpose: To gauge your assessment on the quality and clarity of reporting in the paper

Clarification: Transparency in this context refers to a clear, unambiguous description of the methods used in the research, including experimental procedures, materials, tests and analytical techniques. Think about how easy it is to find all of the information required to perform a close replication of the study. This includes whether or not the study was pre-registered.

Answering the transparency question:

We're asking for this on a scale of 0 to 100. 0 means that the paper is very unclear, not at all transparent in its methods, procedures and/or analyses, 100 means it's perfectly clear and transparent, to the point that another researcher would be able to repeat the methods, procedures and analyses without issue. Again, try to estimate a number and be consistent in the criteria you apply when assessing transparency between papers.

See the tooltip (described below). We don't expect you to thoroughly check everything - that would be time consuming! Just make your best estimate, based on however much of the paper you have read.

Tooltip: We don't expect that you will have thoroughly checked all supplementary materials to ensure the accessibility and reproducibility of relevant materials, data and code. We are asking for your best guess about the level of transparency, based on your rapid assessment of the paper.

Question 4: Replicability

Imagine an independent researcher runs a replication of this original study. What is the probability (0-100%) that a close replication of the central claim would find results consistent with the original paper?

Purpose: The question is asking about a **close replication** of the main claim of the paper. However, if it is easier for you to think about an average replicability of several different claims (because you cannot work out what you think is the central claim, for example), that's ok.

Clarification:

- **A close replication** is a new experiment that follows the methods of the original study with a high degree of similarity, varying only aspects where there is a high degree of confidence that they are not relevant to the research claim. People often use the term direct replication – however no replication is perfectly direct, and we *cannot* describe precisely how any given claim will be replicated. Decisions about how to perform replications are made by a team of researchers that is independent from the repliCATS project. Our best advice is to imagine what kinds of decisions you would face if you were asked to replicate this study, and then to consider the effects of making different choices for these decisions.
- **A successful direct replication** is one that finds a **statistically significant effect** (defined with an alpha of 0.05) that is in the same direction as the original study, using the same statistical technique as the original study.

Specifically, for close replications involving new data collection we would like you to imagine 100 (hypothetical) new replications of the original study, combined to produce a single, overall replication estimate (i.e., a good-faith meta analysis with no publication bias). Assume that all such studies have both a sample size that is at least as large as the original study and high power (90% power to detect an effect 50-75% of the original effect size with alpha=0.05, two sided).

Sometimes it is clear that a close replication involving new data collection is impossible, or infeasible. In these cases, you should think of data analytic replications, in which the central claim is tested against another pre-existing dataset that provides a fair test. Again, imagine 100 datasets analysed with results are combined to produce a single, overall replication estimate.

Answering the replicability question:

In this question, we want you to try and think of reasons why the central claim may or may not replicate. We understand that your thoughts about prior plausibility of this claim is likely to influence your judgement regarding this question. However, we'd like you to try and think more critically of other reasons why this particular study may (or may not) replicate.

Please only use whole integers for this question. Do not use decimal places.

Understanding the three-step format

This question, and some of the following questions, asks you to provide three separate estimates: a lower bound, upper bound and best estimate. Here's how we want you to think of those estimates in regards to replicability. Think about your assessments of *Question 5: Robustness* and *Question 8: Credibility* in a similar manner.

- First, consider all the possible reasons **why a claim is unlikely to successfully replicate**. Use these to provide your estimate of the lowest probability of replication.

- Second, consider the possible reasons **why a claim is likely to successfully replicate**. Use these to provide an estimate of the highest probability of replication.
- Third, consider the balance of evidence. Provide your best estimate of the probability that a study will successfully replicate.

Some things to consider about the three-step format:

- Providing a lower estimate of 0 means you believe a study would never successfully replicate, not even by chance (i.e. you are certain it will not replicate). Providing an upper estimate of 100 means you believe a study would never fail to replicate, not even by chance (i.e. you are certain it will replicate).
- Providing an estimate of 50 means that you believe the weight of evidence is such that it is as likely as not that a study will successfully replicate. If you have low prior knowledge and/or large uncertainty, please use the width of your bounds to reflect this, and still provide your best estimate of the probability of replication.
- Answers **above 50** indicate that you believe it's more likely that the study would replicate than it would not replicate. Answers **below 50** indicate that you believe it's more likely that the study would not replicate than it would replicate. Intervals (the range between your lowest and highest estimate) which extend above and below 50 indicate that you believe there are reasons both for and against the study replicating.

There is evidence that asking you to consider your lower and upper bounds before making your best estimate improves the accuracy of your best estimate. The difference between your upper and lower estimates is intended to reflect your uncertainty about whether that claim's findings would replicate. There's no 'correct' answer here, but we expect that your intervals for those claims you feel most uncertain about will be the widest.

Tooltip: Assume all potential replications are competent, and done in good faith. A close replication is a high-powered, good-faith collection of new data to test the central claims of an original study. Sometimes no such attempt is feasible. In these cases, replication attempts will test central claims with matching analyses on similar, pre-existing datasets. Check training materials for details on close replications and data-analytic replications, including how we define "consistent with the original paper" and high statistical power

Additional considerations on the question of replicability

There are many things you might consider when making your judgement. The IDEA protocol operates well when a diversity of approaches is combined. There is no single 'correct' checklist of things to assess. However, some things you *may* wish to consider include.

- **The statistical data, analyses and results reported within the paper**, including sample size, effect size and p value, if reported. These details are likely to be important for whether a claim replicates - see this document for more information.
- **The experimental design**. Will it be reliable in replication? Are there any signs of Questionable Research Practices e.g. unusual designs where more straightforward tests might have been run but failed? Note that this question is interested in the replicability of the central claim even if the external validity of the design is low.
- **Your prior plausibility** for the paper. Background probabilities are often a major factor. Is this area of research more or less well-understood?
- **Contextual information** about the original study or publication such as where and when the paper was published, and who undertook the original study. Do you have any private or personal knowledge e.g. experience with undertaking similar research, or existing knowledge about the quality of work from a particular source?

Question 5: Robustness

Imagine an independent analyst receives the original data and devises their own means of investigating the central claim of this paper. What is the probability (0-100%) that an alternative analysis would find results consistent with the original paper?

Purpose: To capture your beliefs about the analytic robustness of the main finding

Clarification: The term robustness is used here to represent the stability and reliability of a research finding. It might help to think about it in this way: if 100 analysts received the original data and devised their own means of investigating the central claim, how many would find a statistically significant effect in the same direction as the original?

Answering the robustness question:

This question uses the same 3-point format as *Question 4: Replicability.* To assess the robustness of the central claim of the paper:

- First, consider all the possible reasons **why a claim might not be robust**, i.e. why another researcher using a different analytic approach might find a result that is *inconsistent* with the original claim. Use these to provide your estimate of the lowest probability of a robust finding.
- Second, consider the possible reasons **why a claim would be robust**, i.e. why another researcher using a different analytic approach would find a result that is *entirely consistent or identical* with the original claim. Use these to provide an estimate of the highest probability of a robust finding.
- Third, consider the balance of evidence. Provide your best estimate of the probability that a new analysis of the data would result in a result that is consistent with the original claim.

Tooltip: Assume all potential analyses are competent, and done in good faith. Check training materials for details on how to assess “consistent with the original paper”.

Question 6a: Generalizability (rating)

Going beyond close replications and reanalyses, how well would the main findings of this paper generalize to other ways of researching this question?

Purpose: The question is asking about whether the main claim of the paper would hold up or not under different ways of studying the question.

Clarification: This question is asking about generalizations of the original study, or *conceptual replications*. We want you to consider generalizations across all relevant features, such as the particular instruments or measures used, the sample or population studied, and the time and place of the study.

Answering the generalizability rating question:

This question is asked on a scale of 0 to 100, where 0 means the central claim is not at all generalizable and 100 means it is completely generalizable.

We know that there are many different features of a given study that potentially limit generalizability, and they may have different levels of concern, so it might be tricky to work out a single rating across all of them. Do your best to assess this in whatever way seems best. You might want to imagine 100 (hypothetical) conceptual replications of the original study, each of

which varies a specific aspect of the study design, while holding everything else constant. Across this hypothetical set of conceptual replications, all relevant aspects of the study are varied in turn. How many of these do you estimate would produce a similar finding to the original study?

Question 6b: Generalizability (features)

Please select the feature(s), if any, that you think limit the generalizability of the findings.

Purpose: The question is asking you to list the features of the study that raise substantial generalizability concerns.

Clarification: Select the features for which you definitely have generalizability concerns. Don't select a feature if you simply think that it is *possible* that the study will not generalize over that feature.

Answering the generalizability features question:

You can select more than one feature - select all features you think raise substantial generalizability concerns.

If there is a feature that we have not listed that you think raises substantial generalizability concerns, then select Other and briefly describe the feature in the text box. *Please do not use this text box to discuss **why** you think there are generalizability concerns.* Do that in the comments box for Question 8: Credibility.

Question 7a: Validity (design)

How well-designed is the research presented in this paper to address its aims?

Purpose: The question is asking you to make a judgement about the degree to which the conclusions of the study can be inferred from the reported effect, given the study design.

Clarification: This question focuses on *internal validity*, or the extent to which the study measures what it claims to measure and the methods are suited to address the research aim(s). In a well-designed study, systematic errors and bias can be discounted, such that the outcome can be reliably linked to the (experimental) manipulation or variable of interest. For example, claims about causal relationships among variables need to be warranted by the evidence reported.

Answering the design validity question:

We're asking for this on a scale of 0 to 100. 0 means that the study design is not at all suited to address the research aim(s), 100 means it's perfectly suited to address the research aim(s). Again, try to estimate a number and be consistent in the criteria you apply when assessing the validity of the design between papers.

Tooltip: By "well-designed" we mean that the research has strong methods, e.g., appropriate study design, valid operationalizations and measurement, appropriate sample size, representativeness, experimental control

Question 7b: Validity (analysis)

How appropriate are the statistical tests in this paper?

Purpose: The question is asking about the extent to which a set of statistical inferences, and their underlying assumptions, are appropriate and justified, given the research hypotheses and the (type of) data.

Clarification: This question focuses on a different aspect of *internal validity*, namely on the extent to which the statistical models/tests are appropriate for testing the research hypotheses. For instance, assumptions may have been violated that would render the chosen test(s) inappropriate, or the statistical model of choice may not be appropriate for the type of data.

Answering the analytic validity question:

We're asking for this on a scale of 0 to 100. 0 means that the statistical analyses are not at all appropriate to test the research hypotheses, 100 means they are entirely appropriate to test the research hypotheses. Again, try to estimate a number and be consistent in the criteria you apply when assessing statistical validity between papers.

Tooltip: By "appropriate" we mean that the correct test for the design is done, precautions are taken to minimize risk of bias and error, and assumptions are reasonable.

Question 7c: Validity (conclusions)

How reasonable and well-calibrated are the conclusions drawn from the findings in this paper?

Purpose: The question is asking about the extent to which the paper's conclusions are warranted given the findings.

Clarification: This question relates to the stated interpretation of the findings, whether the conclusions match the evidence presented, and the limitations of the study. Sometimes a paper's conclusion(s) might extend beyond what is indicated by the results reported.

Answering the conclusion validity question:

We're asking for this on a scale of 0 to 100. 0 means that the conclusion is unrelated to the evidence presented, 100 means the conclusion perfectly represents the evidence presented. Again, try to estimate a number and be consistent in the criteria you apply when assessing the validity of the conclusion between papers.

Tooltip: By "well-calibrated" we mean that the conclusions match the evidence presented, without overstepping the mark.

Question 8: Credibility

How would you score the credibility of this paper, overall?

Purpose: The question is asking about how you would assess the overall credibility of a paper, incorporating all of the dimensions that we have asked about so far as well as any other ones you think may be relevant.

Clarification: For this question, we have not provided any specific definition of *Credibility*. We want *you* to determine what you think is credible.

Answering the paper credibility question:

Advice about answering the three separate judgements (lower, upper, best estimate) was given in *Question 4: Replicability* above. Use that advice applied to how you are thinking of *Credibility*.

Everyone is likely to have slightly different understandings about what *Credibility* means. That's ok - there is no one right way of thinking about this. You might want to think about how likely would you be to use this paper if you were in the same field, or apply it to decisions if you were making policy based on it. However, you might have other ways of thinking about it. You may wish to just average across all of the dimensions we have asked about. However, for any given paper, some dimensions may be more important and you may want to weight those dimensions more strongly. And we may have failed to ask about something that you think is important for this paper, so you might include completely different factors in your judgement about *Credibility*.

If you are able to describe how you have thought about overall *Credibility*, we would love you to tell us in the comments box. However, if you cannot articulate your mental model for *Credibility*, that's ok too. Just do your best to make this assessment in whatever way seems best to you.

Tooltip: When giving your lower bound, give the lowest score you would feel comfortable justifying. When giving your upper bound, give the highest score you would feel comfortable justifying. If you feel very unsure about how to score it, make your interval wider.

Comments box.
We have provided a comment box below the question. *Please use the text box for this question to capture all of your thinking about the assessment of this paper.* The credibility of the paper should include all specific dimensions that we have asked about, including both paper-level and evidence-level factors. Collecting all of your reasoning in this one spot may help you to think about what is important for your understanding of *Credibility*. Note that you can navigate to this question - and the text box - at any stage, so you can drop comments in as you go. You don't have to wait until you get to this question before writing your thoughts.

Don't feel that you need to write polished prose here. As long as you can be understood, notes, partial sentences and dot points are fine. However, please do be careful to avoid ambiguities, so that your team members - and the repliCATS project - can make the best use of your comments. For example, if you mention errors or uncertainties, make it clear whether you are referring to the paper itself, the way it is presented on the platform, or how it has been discussed by your team. If you comment on specific results in the paper, make sure you describe which one.

Please do not use your own name or the name of any other participants. Judgements will eventually be made public and we must be able to keep them anonymous. If you want to refer to yourself or a team member, please only use the anonymised screen names.

The placeholder text for this question is:

> *Describe any factors that influenced your judgement about the overall credibility of the paper. Use this textbox for all discussion of any of the specific judgements above, such as the plausibility of the claims, the transparency of the paper, the generalizability of the results or the validity of the research design, or of specific evidence-level results below..*

Evidence-level questions

After answering the questions about the overall paper, we will ask you to assess some specific pieces of evidence within the paper. By doing this, we aim to obtain a more complete evaluation of the paper. You will be asked to consider a variable number of specific results per paper, but no more than ten for any paper. The questions you will be asked are described below. Remember to go back to the comments box for *Question 8: Credibility* if you want to comment on any of these specific pieces of evidence.

Question 9a: Credibility (evidence-level)

Considering the credibility of this particular result alone, it might be the same as the credibility score you gave for the overall paper or it might be different. If different, please change your rating here. We have started here with the credibility score you gave for the overall paper.

Purpose: We understand that different results reported in a given paper vary in credibility. Some results might be highly reliable, while other specific results may be less so. By asking you to separately rate a number of different pieces of evidence, we can get a sense of how much this paper varies.

Clarification: This question relates only to the specific piece of evidence that is listed, both at the top of the question pane and in the left hand sidebar. Some of the statistical information relating to this result has been extracted for you, as well as the location of this specific result in the paper.

Answering the evidence-level credibility question

Think about *Credibility* in the same way that you thought about it in *Question 8: Credibility*. However, here you are considering only the specific result listed, rather than the main claim of the paper or the paper overall.

We have pre-filled this assessment with the overall credibility rating you gave for the paper. If you think that this is an ‘average’ result within the paper then it is ok to just leave the credibility rating as it is. However, if you think that this particular result is more or less credible than the paper overall then please adjust your assessment accordingly.

Note that in Round 2 we will pre-fill your assessment with your Round 1 evidence-level credibility rating, so this is a little bit different. However, we will also remind you of your overall credibility assessment for Round 2, in case your thinking about this has changed between Rounds.

Tooltip: We know that some particular results presented in a paper can vary in terms of their trustworthiness and quality, which is why we’re interested in how credible you find this specific result.

Question 9b: Relevance

How relevant do you think this particular result is to the main conclusion of the paper?

Purpose: We understand that different results reported in a given paper also vary in how important they are. Some results might be central to a paper while others are more peripheral. To fully understand the variability within a paper we need to consider both the credibility of specific results and how important they are.

Clarification: Think about how important this particular piece of evidence is to the main claim of the paper. If this piece of evidence was missing, or turned out to be unreliable, how much would it affect

your confidence in the paper overall?

Answering the relevance question:

This question is asked on a scale of 0 to 100, where 0 means this particular piece of evidence is irrelevant to the main claim, and 100 means it is crucial to your confidence in the main claim. Like other such scales, we just ask you to try to be consistent between papers in how you apply it.

Tooltip: Most papers present some results that are very important or central to the main conclusion, and some results that are more peripheral.We are interested in your evaluation of the relevance of the various results, not in what you think the authors believe.

Question 9c: Replicability (evidence-level)

Particular results presented in a paper can vary in reliability. What is the probability that close replications of this result would find a statistically significant effect in the same direction (0-100%)?

Purpose: For one of the specific results within the paper, we will ask you to assess its replicability. This will allow us to judge whether replicability also varies within papers. It may also allow us to compare different versions of the repliCATS platform.

Clarification: Think about replicability in the same way that you thought about it in Question 4: Replicability. However, here you are considering only the specific result listed, rather than the main claim of the paper or the paper overall.

Answering the evidence-level replicability question:

Like *Question 4: Replicability*, we ask you to provide three separate estimates: a lower bound, upper bound and best estimate.

- First, consider all the possible reasons **why this specific result is unlikely to successfully replicate**. Use these to provide your estimate of the lowest probability of replication.
- Second, consider the possible reasons **why this specific result is likely to successfully replicate**. Use these to provide an estimate of the highest probability of replication.
- Third, consider the balance of evidence. Provide your best estimate of the probability that this specific result will successfully replicate.

Tooltip: You considered the replicability of the central claim of the paper in Q4. Replicability. Again, assume replications are competent, high-powered and done in good faith.

Guidelines for a good Discussion phase

Once you have submitted your Round 1 estimates, you will be able to view estimates and comments made by other participants in your group. It would be particularly helpful for you to examine any comments made by others in response to *Question 8: Credibility*, as we ask people to provide insights on the various dimensions of the paper's credibility in that text box. You can use the repliCATS platform comments feature to react to other participants' judgements, interrogate their reasoning, and ask questions of each other. Please do leave questions and comments for other participants and consider and respond to the questions and comments others might have on yours.

It is important that you do not use participant names on the platform, not even your own.

Reasoning comments will eventually be made public, and we need to keep these anonymous. If you want to refer to other participants, please use the anonymous handle they have been assigned e.g. Koala11.

When everyone's Round 1 judgements have been submitted your group's facilitator will organise a time for you to meet (e.g. via a video-conferencing tool) to directly discuss the paper and ask questions of each other. The Discussion phase in the IDEA protocol is important. It provides an opportunity to resolve differences in interpretation, and to share and examine evidence. In the interest of time and efficiency, your facilitator will focus the discussion on those questions where opinions within the group diverged the most. However, **the purpose of the discussion is not to reach a consensus**, but to investigate the underlying reasons for these (divergent) estimates. The purpose of sharing information in this way is simply that it allows people to reconsider their judgements in light of any new evidence and/or diverging opinions, and the underlying reasons for those opinions.

Ground rules

Some ground rules for the Discussion phase, regardless of whether you are leaving comments on the online platform or discussing directly:

- Respect that the group is composed of a diversity of individuals.
- Consider all perspectives in the group. In synchronous discussion, allow an opportunity for everyone to speak.
- Don't assume everyone has read the same papers or has your skills – explain your reasoning in plain language.
- If someone questions your reasons, they are usually trying to increase their own understanding. Try to provide simple and clear justifications.
- Try to be open-minded about new ideas and evidence.

Tips for a good Discussion phase

The following list may be useful to consider when reviewing and commenting on judgements. You do not have to work through these systematically, but consider which may be relevant:

- What did people believe the claim being made was? Was the paper clear? Did everyone understand the information and terms in the same way? If interpretations of the central claim (or any claim) vary, instead of trying to resolve this, focus on discussing what that means for the credibility of the paper.
- Consider the range of estimates in the group and ask questions about extreme values. What would cause someone to provide a high/low estimate for this question?
- Very wide intervals suggest unconfident responses. Are these based in uncertainties of interpretation or are these participants aware of contradictory evidence?
- Very narrow intervals suggest very confident responses. Do those participants have extra information?
- It's ok if you don't have good evidence for your beliefs – please feel free to state this.
- If you have changed your mind since your Round 1 estimates it's good to share this. Actively entertaining counterfactual evidence and beliefs improves your judgements.
- If you disagree with the group that is fine. Please state your true belief when completing Round 2 estimates. This represents the true uncertainty regarding the question, and it should be captured.
- Consider raising counterarguments, not to be a nuisance, but to make sure the group considers the full range of evidence.
- As a group, avoid getting bogged down in details that are not crucial to answering the questions, or in trying to resolve differences in interpretation. Focus on sharing your reasons, not on convincing others of specific question answers.

Guidelines for Round 2

Entering your Round 2 estimates

You may wish to update your estimates in real-time, as you and your group go through and discuss the paper and the evidence-level claims. Alternatively, you may wish to update after the discussion. However, we do advise that you enter your Round 2 estimates as soon as possible following the discussion, while the information that was shared is still fresh.

Whether or not you want to update your estimates is entirely up to you. In some instances, your views and opinions might not have shifted after discussion, but perhaps not. Either decision is absolutely fine and provides useful information. Remember, your Round 2 estimates are private judgements, just like your Round 1 estimates, so there is no need to worry about what others might do or think about your judgements.

Help and FAQs

We understand the task assigned to you is not easy, here are some tips to help you through.

I feel unqualified to answer make these assessments

There will be papers which you feel you are unqualified to assess. This is natural, the task is difficult. We ask you to please attempt all papers assigned. We do not expect you to have specific expertise in every paper you assess. We expect that both experts and non-experts contribute to good judgements. People who consider themselves 'outsiders' may notice details about the paper which more experienced members of the group overlook. However, if you genuinely cannot understand a paper, it will not be a satisfying experience to assess it. Thus, our standard is that you *feel comfortable* interpreting a paper, **not** that you *feel expert* in the subject. Remember:

- You can adjust your upper and lower bounds to express your uncertainty.
- You can draw on thoughts and opinions of colleagues (as long as they are not participants assessing the same papers), or even additional resources.
- You can also justify your responses and any evidence or questions you have about the claim in the comments box to *Question 8: Credibility*.
- In Round 2 you'll be able to draw on the knowledge of the other participants and update your response in light of the Discussion phase.
- If you feel you really don't understand the paper, then please note this and provide your best interpretation of the paper's meaning. We can discuss these interpretations in the Discussion phase.

I'd like to access some training materials and practice questions

We have also developed some training which might help you to better assess the claims. This includes a downloadable training document that covers statistical concepts and some background information on questionable research practices and replication rates in previous studies. This training document can be downloaded here. We will also send you an interactive quiz and example claims to practice on. Including both the document and quiz, the training takes about one hour to complete.

I don't understand what is meant by 'replication' and other terms

A list of terms can be found in our glossary. If there is a term missing, please notify us.

I am involved in a replication study for a claim that I have been assigned

In the unlikely event that you are involved in the replication of one or more of the claims assigned, then please do not assess that claim, and please do not reveal which claim you are involved in replicating. Tell the workshop co-ordinator, or another repliCATS team member.

Will my data be made publicly available? Can I withdraw it?

At the end of this project, data will be made publicly accessible in an anonymized format. It is for this reason that you ***must not use the name of any participant including your own.*** While you can end your participation at any time, we may not be able to remove data previously submitted to us.

Code of conduct

The repliCATS project has a strict code of conduct. Please report any suspected breach.

Who can I contact about…

- ❏ *General repliCATS questions*: Mel Ross repliCATS-contact@unimelb.edu.au
- ❏ *Trouble logging in:* Mel Ross repliCATS-contact@unimelb.edu.au
- ❏ *Other platform troubleshooting:* Report through platform
- ❏ *General questions about assessing claims:* Refer to participant materials
  If still unresolved, send query to repliCATS-contact@unimelb.edu.au
- ❏ *Code of conduct breaches:* Report in platform or to Fallon Mody
  fallon.mody@unimelb.edu.au

Who can I contact if I have concerns about the project itself?

This research project has been approved by the Human Research Ethics Committee of The University of Melbourne. If you have any concerns or complaints about the conduct of this research project, which you do not wish to discuss with the research team, you should contact the Manager, Human Research Ethics, Research Ethics and Integrity, University of Melbourne, VIC 3010. Tel: +61 3 8344 2073 or Email: humanethics-complaints@unimelb.edu.au. All complaints will be treated confidentially. In any correspondence please provide the name of the research team (provided above) and ethics ID number of the research project 1853445.