Peer Evaluation Procedure

The peer evaluation procedure described below follows our preregistered protocol. All minor deviations are listed in the Deviations from Preregistration document in the Supplementary materials.

In addition to completing a re-analyses, co-analysts who signed up to the Multi100 could opt-in to serve as a peer evaluator on the project. That is, the co-analyst who responded ‘Yes’ when asked “Are you interested in serving as an evaluator for this project?” were later approached to serve as peer evaluators. The role of a peer evaluator was to check the plausibility and sensibility of an analysis based on a summary of the analysis submitted by the analyst. In order to successfully evaluate a given re-analysis, peer evaluators were provided with instructions and a summary of the co-analysts analysis (i.e., responses to their Task 1 and Task 2 post-analysis survey questions). A template of instructions is provided in the Figure below.

**Figure**

*Peer evaluation task template sent out to all evaluators. Square brackets indicate variable information that is specific to the re-analysis being evaluated.*

[INSERT FIGURE HERE]

## Main outputs of the peer evaluation

Re-analyses were evaluated on five key metrics.

First, peer evaluators judged whether the analysis pipeline of task 1 was acceptable. That is, they judged whether it is within the variations that could be considered appropriate by the scientific community in addressing the underlying research question. Each re-analysis pipeline was rated as either (1) Unacceptable, (2) Acceptable but low quality, (3) Acceptable and medium quality, or (4) Acceptable and high quality. In cases where the analysis pipeline was deemed to be unacceptable, evaluators provided their reasoning via an open text box.

Second, peer evaluators judged whether the conclusion provided in task 1 adequately followed from the results of the analysis. Each conclusion was rated as either (1) It adequately follows from the results of the analysis, or (2) It does not follow adequately from the results of the analysis. In cases where the conclusion was judged not to follow adequately from the results, evaluators provided their reasoning via an open text box.

Third, peer evaluators judged whether the analyst’s categorization of the task 1 result is adequate. For example, regarding an analyst who has claimed that the results of their analysis show evidence in favor of the original effect/relationship, the evaluator considered whether this judgment is adequate. Each categorization was rated as either (1) Adequate, or (2) Inadequate. Given that the analyst’s categorization of the results is tied to their conclusion, there was no open text box provided for inadequate ratings.

Fourth, peer evaluators judged whether the analysis pipeline of task 2 was acceptable. That is, they judged whether it is within the variations that could be considered appropriate by the scientific community in addressing the underlying research question. Each re-analysis pipeline was rated as either (1) Unacceptable, (2) Acceptable but low quality, (3) Acceptable and medium quality, (4) Acceptable and high quality, or (5) Incomplete or missing analysis. In cases where the analysis pipeline was deemed to be unacceptable, evaluators provided their reasoning via an open text box.

Finally, peer evaluators could optionally complete a code reproducibility check. They were asked whether any mismatches were found between the executed code and the reported results. For each analysis, the evaluator indicated either (1) I didn’t try to execute it, (2) I tried but didn’t manage to execute it, (3) I executed it and I found no mismatches, or (4) I executed it and I found mismatches. In cases where mismatches were found, evaluators described the nature of these mismatches via an open text box.

Accordingly, the task 1 analysis pipeline was rated as ‘Unacceptable’ in 8 cases, the task 1 conclusion was judged not to follow adequately from the results in 27 cases, the task 1 self-categorization of the result was rated as ‘inadequate’ in 38 cases, the task 2 analysis pipeline was rated as ‘unacceptable’ in 18 cases while the task 2 analysis pipeline was judged as ‘incomplete or missing’ in 21 cases, while the code reproducibility checks revealed 19 mismatches.

## Review of the peer evaluation reports

To identify potential errors or misunderstandings in the peer evaluations, each issue raised (above) by a peer evaluator was reviewed by a member of the expert panel who considered the information provided by the peer evaluator and where necessary, contrasted it with the information provided by the co-analyst. For each issue, the panel member reviewed the evaluators’ initial categorization and their reasoning. Note, that our aim in the project was to explore the sensitivity of analytical results to the analytical choices of the co-analysts. So during the process of peer evaluation, our goal was not to ensure that each analysis pipeline consisted of the most ideal steps from every possible perspective but to ensure that the steps of the analyses were within the variations that could be considered appropriate by the scientific community in addressing the given analytical task. For that reason, during the review of the peer evaluations, the expert panel member left the ratings of the peer evaluators ‘Unacceptable’ only if the analyst made one or more mistakes that could be objectively judged as incorrect. For all the other cases where the peer-evaluator categorized the analysis pipeline ‘Unacceptable’ based on non-objective reasoning (e.g., not adding control variables or controlling for another variable), the expert panel member adjusted the rating from ‘Unacceptable’ to ‘Acceptable but low quality’. We aimed to remove re-analyses from the dataset where it was judged as ‘Unacceptable’.

As a consequence of the full peer evaluation review, one analysis was rejected. What follows is a summary of revisions made to peer evaluator’s initial ratings as an outcome of the peer evaluation review.

Following the task 1 analysis pipeline review, ratings of ‘(1) Unacceptable’ (n = 7) were revised to ‘(2) Acceptable but low quality’. Following the task 1 conclusion review, ratings of ‘(2) It does not follow adequately from the results of the analysis’ (n = 25) were revised to ‘(1) It follows adequately from the results of the analysis’.

Following the task 1 categorization review, ratings of ‘(2) Inadequate’ (n = 36) were revised to ‘Adequate’. In many cases, evaluators made their judgment of ‘inadequate’ on the basis of their task 1 conclusion rating. Put simply, evaluators often considered the categorization of results to be inadequate when they also judged that the conclusion does not follow from the results. It was often the case then, that verifying the legitimacy of the task 1 conclusion also verified the legitimacy of the task 1 categorization. The reasoning of the expert panel on a case-by-case basis can be found in the review supplement.

Following the task 2 analysis pipeline review, ratings of ‘(1) Unacceptable’ (n = 17) were revised to ‘(2) Acceptable but low quality’. All initial ratings of ‘(5) Incomplete or missing analysis’ (n = 21) were also revised. Many of these ratings were made simply because the re-analysts task 1 submission also satisfied the requirements of task 2 (i.e., the paper-specific instructions given in task 2 had already been adhered to in task 1), and as a result, no further analysis was needed. For each case, the panel verified that the analyst had reported their test statistic appropriately in the task 2 survey response, and that their analysis files had been uploaded to the OSF as requested.

Finally, there were no changes made to initial ratings following the code mismatches review. In the cases where evaluators reported ‘(4) I executed it and found mismatches’ (n = 19), the panel verified that the mismatches did not have a meaningful impact on the re-analyst’s reported conclusion, categorization, or effect size.

The issues raised by the peer evaluators and the decisions of the expert panel are documented in full on the ‘Peer Evaluation: Review and Decisions’ supplement. This document also contains the panel’s reasoning in each case.

## Peer Evaluation Results

See Supplement 2 - Results document.

## Resulting actions

Re-analyses: Rejected re-analyses were excluded from overall data analyses. Those co-analysts with no accepted analyses were not co-authors on the resulting publication unless they earned it by completing peer evaluations.