nature portfolio

https://doi.org/10.1038/s41586-025-09844-9

## Supplementary information

# Investigating the analytical robustness of the social and behavioural sciences

In the format provided by the authors and unedited

# Supplementary information

## Investigating the analytical robustness of the social and behavioural sciences

### Table of Contents

## Additional Results

### General descriptives

Responding to our recruitment call, 1141 researchers signed up to express initial interest in participating in our study. Of these, 459 ultimately analysed at least one dataset and submitted their work by the deadline or an extended deadline.

Throughout the project, 509 re-analyses were submitted. This number is higher than the number of re-analysts, as some re-analysts volunteered to analyse more than one dataset.

Out of the submitted analyses, one was omitted from the summary analysis as it failed the peer evaluation, and an additional four analyses were excluded due to incomplete responses.

As a result, we ended up with 504 re-analyses submitted by 457 re-analysts.

Although we invited more than 5 re-analysts to work on each of the 100 studies, due to dropouts and peer evaluation exclusions, the final number of completed analyses ranged from 3 to 7. Table S1 shows the distribution of the number of analyses for individual studies.

**Table S1 | The Distribution of the Number of Analyses for Studies**

| Number of Completed Analyses | Number of Studies |
|---|---|
| 3 | 1 |
| 4 | 13 |
| 5 | 69 |
| 6 | 15 |
| 7 | 2 |

### Demographics of the re-analysts

Out of all the re-analysts who submitted their work by the deadline, there were 23 professors, 41 associate professors, 105 assistant professors, 107 post-doctoral researchers, 122 doctoral students, and 59 from other academic/research positions.

The gender distribution of the re-analysts was as follows: 117 females, 332 males, 1 other, and 7 did not want to respond to this question.

The age distribution of the re-analysts was as follows: 375 young adults (<39 years); 81 middle-aged adults (40-59 years); and no older adults (>60 years).

Regarding the highest level of education, one re-analyst reported a high school diploma or equivalent, 18 re-analysts had a Bachelor's degree or equivalent, 135 had a Master's degree or equivalent, and 303 had a Doctoral degree or equivalent. In case the analysts completed more than one re-analysis and advanced in their studies by the time of their second analysis, we kept only their first response for this comparison.

Regarding the continents, one re-analyst was from Africa, 27 were from Asia, 15 from Oceania, 296 from Europe, 112 from North America, and six were from South America.

The median duration of experience with data analysis was eight years among our re-analysts.

We asked our re-analysts how regularly they perform data analysis. The most frequent category was 2-3 times a week.

We also asked them to rate their level of expertise in data analysis on a scale of 1 (Beginner) to 10 (Expert). The most prevalent answer was 8.

In 8.13% (41 out of 504) of the cases, the re-analysts were familiar with the paper that the provided dataset belongs to before beginning their work on the project.

All re-analysts reported that they had not communicated the details of their analysis with other re-analysts working with the same dataset.

We asked the re-analysts which programming language, software, or tool they used in their data analysis during Tasks 1 and 2. R (62.53%), STATA (16.86%), and SPSS (7.02%) were the most popular responses.

We asked the re-analysts which discipline is closest to their research area. Table S2 summarises the distribution of their disciplinary orientations. The largest subgroups of re-analysts were from Psychology and Economics.

**Table S2. The Distribution of Re-analysts' Disciplinary Orientation**

| Discipline | Count | Percentage |
|---|---|---|
| Psychology | 264 | 57.77 |
| Economics | 74 | 16.19 |
| Political Science | 34 | 7.44 |
| Business Studies | 27 | 5.91 |
| Sociology | 19 | 4.16 |
| Computer Science/Statistics/Data Science | 16 | 3.50 |
| Public Policy | 3 | 0.66 |
| Anthropology | 1 | 0.22 |
| International Relations | 1 | 0.22 |
| Other | 18 | 3.94 |

Note: Whenever the respondents provided more than one field, we only kept their first responses.

## Descriptives of the statistical analyses

The primary difference between Task 2 and Task 1 was that the re-analysts were given specific constraints on their analysis, which focused on a single result from the original study (see Methods for more details).

In Task 2, when we asked the re-analysts to present one main statistical result, in 97.62% of the analyses (492 out of 504), the conclusion was based on a *p*-value. A Bayes Factor was chosen in 2.38% of the cases (12 out of 504).

For 47.82% (241 out of 504) of the analyses, the re-analysts reported having to make additional calculations in Task 2 compared to Task 1. In the remaining 52.18% (263 out of 504) of the cases, the re-analysts indicated that they could conduct the same analyses as in Task 1 and meet the requirements of the task instructions.

In Task 2, 12.7% of the results (64 out of 504) were in the opposite direction from those claimed by the original study, disregarding whether the effect was conclusive/significant.

The re-analysts were asked to estimate their time spent performing Task 1 and Task 2 together. The median value of their response was 6 hours.

## Peer evaluation

### *Descriptives of the peer evaluators*

Most peer evaluators have many years of experience conducting statistical analysis, perform data analysis regularly, and rate themselves close to expert level in data analysis.

### *Descriptives of peer evaluations*

In total, we received 490 peer evaluation reports. One peer evaluation was removed because the analyst's ID was not provided, and as such, we could not verify with certainty which re-analysis was being evaluated, leaving us with a total of 489 peer evaluation reports for 73 different papers. After the panel members had reviewed the peer evaluations (see 'Peer Evaluation: Review and Decisions' for all decisions and reasoning behind each case), the final result of the peer evaluation was as follows:

At the end of the peer evaluation process, one analysis contained an unacceptable analysis pipeline due to quality concerns. Therefore, we removed this single analysis from our results. For the remaining analyses, it was determined that all Task 1 and Task 2 analysis pipelines were acceptable. Furthermore, all remaining Task 1 conclusions were considered to follow from the results accurately, and the analysts' self-categorisation of the results was deemed adequate.

204 analytical reproducibility checks were successfully conducted, and mismatches were identified in 19 analyses. In all of these cases, we verified that the mismatches did not have a meaningful impact on the reported conclusion, categorisation, or effect size.

## Robustness analyses

### *Robustness of the conclusions by the level of confidence in the suitability of the analysis*

Table S3 shows the percentage of same conclusion, no effect/inconclusive, and opposite effect of the re-analyses by the analyst's level of confidence with the suitability of the analysis.

**Table S3. Robustness of the Conclusions by the Level of Confidence with the Suitability of the Analysis**

| Confidence rating | Direction of the conclusion | Count | Percentage |
|---|---|---|---|
| 1 Not confident at all | Same conclusion | 1 / 3 | 33% |
| 1 Not confident at all | No effect/inconclusive | 2 / 3 | 67% |
| 1 Not confident at all | Opposite effect | 0 / 3 | 0% |
| 2 | Same conclusion | 11 / 15 | 73% |
| 2 | No effect/inconclusive | 4 / 15 | 27% |
| 2 | Opposite effect | 0 / 15 | 0% |
| 3 | Same conclusion | 43 / 81 | 53% |
| 3 | No effect/inconclusive | 35 / 81 | 43% |
| 3 | Opposite effect | 3 / 81 | 4% |
| 4 | Same conclusion | 165 / 228 | 72% |
| 4 | No effect/inconclusive | 57 / 228 | 25% |
| 4 | Opposite effect | 6 / 228 | 3% |
| 5 Very confident | Same conclusion | 151 / 177 | 85% |
| 5 Very confident | No effect/inconclusive | 24 / 177 | 14% |
| 5 Very confident | Opposite effect | 2 / 177 | 1% |

### *Robustness of the statistical results by the level of confidence in the suitability of the analysis*

Here (Table S4), we were interested to see whether these results show a different pattern when inspecting them as a function of the evaluators' level of confidence in the suitability of the analysis.

**Table S4. Robustness of the Statistical Results by the Level of Confidence with the Suitability of the Analysis**

| Confidence rating | Is the estimate within the tolerance region? | Count | Percentage |
|---|---|---|---|
| 1 Not confident at all | Yes | 2 / 3 | 66.67% |
| 1 Not confident at all | No | 1 / 3 | 33.33% |
| 2 | Yes | 5 / 13 | 38.46% |
| 2 | No | 8 / 13 | 61.54% |
| 3 | Yes | 14 / 68 | 20.59% |
| 3 | No | 54 / 68 | 79.41% |
| 4 | Yes | 51 / 192 | 26.56% |
| 4 | No | 141 / 192 | 73.44% |
| 5 Very confident | Yes | 59 / 141 | 41.84% |
| 5 Very confident | No | 82 / 141 | 58.16% |

***Robustness of the statistical results with alternative tolerance regions***

Even with a 4 times broader tolerance region (+/- 0.5 Cohen's *d*), around 80% of the studies still show results outside of this region, and around half of the individual reanalysis effect sizes are outside of this region (Extended Data Fig. 4a).

Alternatively, we could define the tolerance region as a percentage of the given effect size. As an additional robustness test, we varied the tolerance region between +/-5% and +/-20%. Extended Data Fig. 4b shows that there was barely any difference regarding the percentage of robust studies.

**Generalised marginal effects**

While Cohen's *d* has the advantage of being easy to compute and comparable across different analyses, Kümpel and Hoffmann[26] recently proposed the concept of generalised marginal effects (gMEs), an effect size metric that is both formally applicable and comparable across different statistical models. When standardised, the value of gMEs is equal to the value of Cohen's *d,* where the latter effect size measure is strictly applicable. Otherwise, gME-values are intuitive to interpret and communicate, as they give the average expected change of the target variable. We had not originally planned to calculate standardised gMEs, and, accordingly, did not collect all required analysis outputs to compute them across the board. Still, we calculated gMEs for a sample of our studies ($n = 4$) to showcase their potential for future multi-analyst studies (Fig. S1).

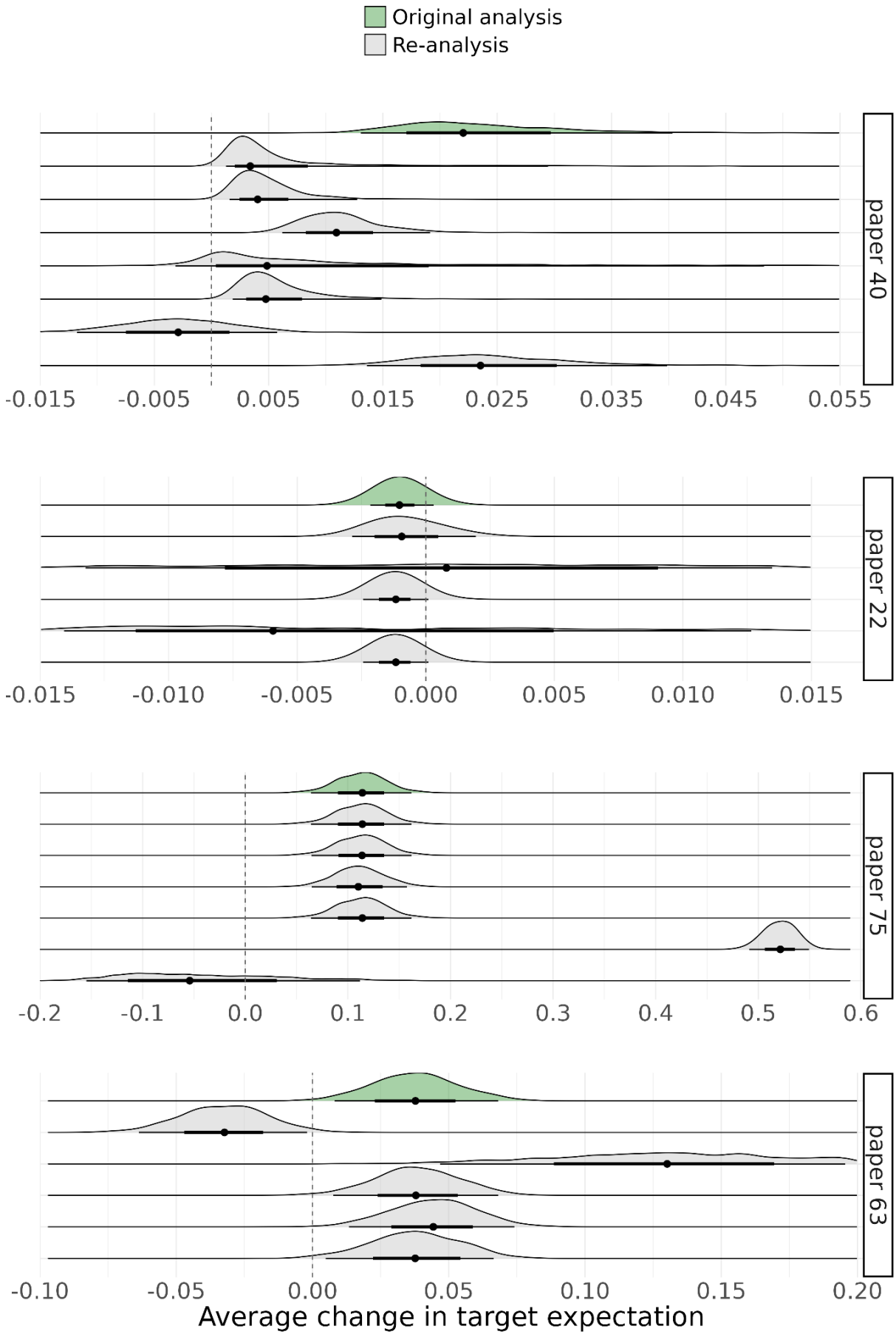
Original analysis
Re-analysis
paper 40
-0.015
-0.005
0.005
0.015
0.025
0.035
0.045
0.055
paper 22
-0.015
-0.010
-0.005
0.000
0.005
0.010
0.015
paper 75
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
paper 63
-0.10
-0.05
0.00
0.05
0.10
0.15
0.20
Average change in target expectation

**Fig. S1 | gMEs for a sample of the studies**
For each original and re-analysis of papers 22, 40, 63, and 75, the figure shows a forest-density plot of non-standardised gME values, as defined by Kümpel and Hoffmann[26]. Specifically, the black dots give the point estimates of the average change in target expectation attributed to the regressor of interest by each analysis, whereas the thicker and thinner lines depict the 0.66 and 0.95 quantiles of the corresponding densities. Study numbers correspond to studies listed in Table S1.

## Analyses requested during the review process

### *Robustness by analysis method*

In the following, we explore the robustness of the statistical results separated by whether the re-analyst used the same or different analysis method as the original study, displayed separately for experimental and observational designs.

**Table S5. | The Number and percentage of matching analyses by study design type**

| Statistical test family | *n* | *N* | Percentage |
|---|---|---|---|
| Experimental | | | |
| Match | 93 | 147 | 63.27 |
| No match | 54 | 147 | 36.73 |
| Observational | | | |
| Match | 141 | 249 | 56.63 |
| No match | 108 | 249 | 43.37 |

**Table S6. | The Number and percentage of matching analyses by study design type and their relationship to the robustness of the re-analysis effect sizes**

| Within tolerance region | *n* | *N* | Percentage |
|---|---|---|---|
| Experimental - Match | | | |
| Yes | 44 | 93 | 47.31 |
| No | 49 | 93 | 52.69 |
| Experimental - No match | | | |
| Yes | 21 | 54 | 38.89 |
| No | 33 | 54 | 61.11 |
| Observational - Match | | | |
| Yes | 47 | 141 | 33.33 |
| No | 94 | 141 | 66.67 |
| Observational - No match | | | |
| Yes | 22 | 108 | 20.37 |
| No | 86 | 108 | 79.63 |

***Robustness of the statistical results separated by matches and non-matches between the discipline of the re-analyst and the original study***

In this additional analysis, we matched the discipline of the original paper with the background of the re-analysts. For 404 re-analyses (from the fields of economics, political science, psychology, and sociology) where we could unambiguously pair the discipline of the study and the analysts' background, we found a match in 46% of the cases. We recalculated the main results presented in the abstract for each group. We found that 34% (50 out of 147) of the re-analysis effect sizes were within the .05 tolerance region for matching disciplines, while this value was 29% (51 out of 173) for non-matching disciplines. For the matching disciplines, 74% (135 out of 183) of the analyses reached the same conclusion, 24% (43 out of 183) were inconclusive, and 3% (5 out of 183) were in the opposite direction from the original claim. In the non-matching disciplines, 70% (153 out of 219) of the analyses reached the same conclusion as the original study, 29% (64 out of 219) were inconclusive, and 1% (2 out of 219) drew the opposite conclusion. Overall, these results show very similar patterns for matching and non-matching disciplines, suggesting that the main results are not highly dependent on a match between a paper's field and an analyst's background. We present these results in Tables S7, S8, S9, and Extended Data Fig. 4a.

**Table S7. | The matches and nonmatches between the discipline of the re-analyst and the original study**

| Discipline of original paper and re-analyst | *n* | *N* | Percentage |
|---|---|---|---|
| Match | 183 | 402 | 45.52 |
| No match | 219 | 402 | 54.48 |

**Table S8. | The robustness of the conclusions by matches and nonmatches between the discipline of the re-analyst and the original study**

| Within tolerance region | *n* | *N* | Percentage |
|---|---|---|---|
| Match | | | |
| Yes | 50 | 147 | 34.01 |
| No | 97 | 147 | 65.99 |
| No match | | | |
| Yes | 51 | 173 | 29.48 |
| No | 122 | 173 | 70.52 |

**Table S9. | The robustness of the statistical results by matches and nonmatches between the discipline of the re-analyst and the original study**

| Discipline of original paper and re-analyst | Categorisation | *n* | *N* | Percentage |
|---|---|---|---|---|
| Match | Same conclusion | 135 | 183 | 73.77 |
| Match | No effect/inconclusive | 43 | 183 | 23.50 |
| Match | Opposite effect | 5 | 183 | 2.73 |
| No match | Same conclusion | 153 | 219 | 69.86 |
| No match | No effect/inconclusive | 64 | 219 | 29.22 |
| No match | Opposite effect | 2 | 219 | 0.91 |

***Heterogeneity ratio of the re-analysed studies***

To calculate the relation between the variability over the re-analyses and the sampling variability of the original study effect size estimates, the heterogeneity ratio, we divided the standard deviation across effect size estimates in Cohen's *d* (as a proxy of between-analysis variability) by the mean standard error of the estimates in the same units (as a proxy of the sampling variation) using the method suggested by Huntington-Klein et al. (2021) (also see Holzmeister et al., 2024). For the calculation, the results of 417 re-analyses were applicable. The median heterogeneity ratio across all papers was 2.03 (IQR = 4.77), which indicates that, on average, the variability due to analytical heterogeneity is about twice as large as the sampling variation. The distribution of the heterogeneity ratios is shown in Fig. S5b.

This calculation was not planned in advance, and the required statistical input values can greatly reflect the re-analysts' reporting preferences. Therefore, we regard these results as rough estimates and do not discuss them in the manuscript.

Holzmeister, F. *et al.* Heterogeneity in effect size estimates. *Proceedings of the National Academy of Sciences.* 121, e2403490121 (2024).

Huntington-Klein, N. et al. The influence of hidden researcher decisions in applied microeconomics. Econ. Inq. 59, 944–960 (2021).

***Robustness results only when the original results were analytically reproduced***

For 78 out of the 100 studies, the COS team could reproduce the original results. For the 78 papers where the original analysis could be reproduced, 36% (118 out of 327) of the present re-analyses yielded the same results (within a tolerance region of +/- 0.05 Cohen's d). Regarding the conclusions drawn, 72% (286 out of 395) of the re-analyses arrived at the same conclusion, 25% (98 out of 395) were inconclusive, and 3% (11 out of 395) of the re-analyses arrived at the opposite conclusion as the original study. For the 22 studies where the original results were not reproducible, 23% (16 out of 69) of the re-analyses yielded the same result within a tolerance region of +/- 0.05 Cohen's d. 78% (85 out of 109) of the re-analyses arrived at the same conclusion, and 22% (24 out of 109) were inconclusive. No re-analysts arrived at the opposite conclusion as the original study.

Please note that those studies categorised as 'not being computationally reproduced' when added to the sample showed only minor differences in the output. Further, 'not being computationally reproduced' covers cases such as the validation failing due to technical limitations, or the analysis code not being available.

***Robustness results without outlier effect sizes***

Without any outliers (Cohen's d >= 3 and <= -3), we have 433 (out of 504) re-analysis effect sizes. 35% of the re-analysis effect sizes are within a tolerance region (+/- 0.05 Cohen's d), which is almost the same as our overall finding. Although we report these results here, we found the analyses of d > |3| valid and saw no reason to drop these analyses from the study.

***Robustness results only for studies with openly available data***

We checked each study to determine whether it had open data. It is worth noting that data openness is a spectrum. We encountered several instances where data were technically open but still unavailable due to expired links, closed-access articles, defunct pages, etc. We settled on an

operationalisation where only a few clicks should be enough to get the data. Otherwise, we labelled the data as non-open. Out of the 100 studies, 47 had available data upon further examination, and 53 did not. Where the original data were available, 24% (43 out of 182) of the re-analysts arrived at the same result as the original study (within a tolerance region of +/- 0.05 Cohen's d). In contrast, this value was 37% (88 out of 235) for those papers for which we could not easily locate the original data. Regarding the conclusions drawn, where the data were available, 75% (173 out of 231) of analyses were reported to arrive at the same conclusion as in the original investigation; 23% (54 out of 231) came to no effects/inconclusive result, and 2% (4 out of 231) to the opposite effect as in the original investigation. These values were 73% (same; 198 out of 273), 24% (inconclusive; 68 out of 273), and 2% (opposite; 7 out of 273), where the data were not available. We include the caveats that these calculations are very rough and the connection between data sharing and robust analytic procedures is, to some degree, speculative and indirect. Furthermore, this can be influenced by various external factors (e.g., data sharing being more common in certain fields than others).

## Additional Details of Method

### Procedures

#### *Re-analyst recruitment*

Our preregistered aim was to have at least five independent re-analyses carried out for each of the 100 selected studies (Fig. S16). Our choice of 5 analyses per study was led by practical considerations, as we judged that recruiting 500 analysts for a project is the limit of our capacity.

Participation in the project was advertised on social media, at conferences, in mailing lists (e.g., SCORE collaborator list), via personal networks, and in research newsletters. As a response to our recruitment call, 1141 researchers signed up to participate in our study. Out of these volunteers, 459 signed up to analyse at least one dataset and submitted their work by the deadline or an extended deadline.

### Project contributors

*Lead team:* The project was coordinated by a lead team (consisting of Balazs Aczel, Barnabas Szaszi, Harry Clelland, Livia Kosa, Zoltan Torma, Felix Holzmeister, Marton Kovacs, and Gustav Nilsonne). The lead team was responsible for the development of the research methodology, preregistration of the project, overall analysis of the results, and preparation of a manuscript for publication. Furthermore, the lead team provided the materials to re-analysts and peer evaluators; and communicated with the project management team, the expert panel, re-analysts, and peer evaluators to ensure that the project proceeded as intended.

*Project management team:* The project management team consisted of the SCORE team of the Center for Open Science and recruited research assistants. The project management team provided financial support, oversaw the legal and ethical aspects of the project, provided the infrastructure

and support for data management, and supported the use of materials adopted from other SCORE projects.

*Expert panel:* A group of experts who have previously conducted multi-analyst studies and/or are experts in relevant methodology were invited to participate in the project as members of an expert panel. The panel's task was to oversee the research plan and remain available to comment on methodological questions throughout the project. The list of expert panel members is available at https://osf.io/j3a9k.

*Re-analysts:* Analysts who independently analysed the target datasets.

*Peer evaluators:* Peer evaluators were also re-analysts who were asked to evaluate the completed analyses.

## Materials

### *Study and claim selection*

After selecting 100 studies from our collection, we selected one empirical claim from each (see Method). We provided the re-analysts with the claims to test on the original datasets, but we did not give them specific research questions. Instead, we used the selected claims to focus the analysts on an underlying research question. We decided to follow this approach because the original papers rarely contained a fully specified research question, and we judged that any attempt to translate the extracted claims into research questions would carry the risk of influencing the analysts based on our own interpretation.

## Peer evaluations

### *Peer evaluators*

When volunteering to be a re-analyst in the project, researchers could indicate whether they would be willing to serve as peer evaluators as well. They were informed that peer evaluators could become co-authors of the resulting article and that they would be remunerated for their efforts. Peer evaluators were asked to deliver five evaluations by a predefined due date and will be paid a flat fee of $10 per evaluation as an incentive to comply with the agreement. 8 peer evaluators evaluated more than 5 ( 6-10) analyses and were paid accordingly.

We initially aimed to recruit at least two peer evaluations for each re-analysis. Therefore, our aim was to reach a total of 100 x 5 x 2 = 1,000 evaluations. In reality, this plan turned out to be overly ambitious due to the very labour-intensive coordination and assessment work. Therefore, after completing an initial 507 peer evaluations, we judged that our sample could provide us with a rough estimate of the potentially unacceptable analyses, and since we found this value relatively low (see Results), we ceased to continue with further peer evaluations.

**Peer Evaluation Procedure**

The peer evaluation procedure described below follows our preregistered protocol. Deviations are listed in the 'Deviations from Preregistration' document.

In addition to completing a re-analysis, re-analysts who signed up for the Multi100 could opt in to serve as peer evaluators on the project. That is, the re-analyst who responded 'Yes' when asked, "Are you interested in serving as an evaluator for this project?" was later approached to serve as a peer evaluator. The role of a peer evaluator was to check the plausibility and legitimacy 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 re-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.

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

Re-analysis Summary Report for Peer-evaluation

*Paper ID:* [Paper ID]
*Re-analyst's ID:* [Analyst ID]

**Your task**

As a peer-evaluator, for this paper, you are asked to judge whether the analyst's applied analytical choices for Task 1 and Task 2 are acceptable. By acceptable, we mean that the analysis pipeline is within the variations that could be considered appropriate by the scientific community in addressing the given analytical task. In addition, for Task 1, we ask you to judge whether the reported conclusion adequately follows from the results of the analysis; whether the re-analyst's self-categorization of the result is adequate. Although, you are not required, it would be appreciated, if you could execute the accompanying analysis code to check any mismatch between the results of the analysis and the reported results.

*For both tasks, you should provide your evaluation via this survey:* [LINK]

**The re-analyst's task**

Please assess the following claim of the original authors:
[Original Claim]

**The corresponding article**

Here you can find the original article: [Link to Original Paper]
and the original data: [Link to Original Data]

They are provided only as context of the re-analysis.

**In Task 1, the analyst was asked to conduct the analysis without any restrictions. Here, the analyst reported the following steps and results of the analysis:**
[Co-Analyst's Response via Task 1 Post-Analysis Survey]

**The analyst's conclusion:**
[Co-Analyst's Response via Task 1 Post-Analysis Survey]

**The analyst's categorization of the result:**
[Co-Analyst's Response via Task 1 Post-Analysis Survey]

**In Task 2, the analysts received the following instruction to conduct the analysis:**
[Paper-specific Instructions Given]

**Here, the analyst reported the following steps and results of the analysis:**
[Co-Analyst's Response via Task 2 Post-Analysis Survey]

**Analysis language/software/code (optional to check):**
Task 1: [Co-Analyst's Response via Task 2 Post-Analysis Survey]
Task 2: [Co-Analyst's Response via Task 2 Post-Analysis Survey]

**Analysis codes:** [Link to Analyst's OSF Page]

### 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 unacceptable, evaluators provided their reasoning via an open text box.

Second, peer evaluators judged whether the conclusion provided in Task 1 adequately followed 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 categorisation of the Task 1 result was adequate. For example, regarding an analyst who has claimed that the results of their analysis show evidence in favour of the original effect/relationship, the evaluator considered whether this judgment is adequate. Each categorisation was rated as either (1) Adequate, or (2) Inadequate.

Given that the analyst's categorisation 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 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, the Task 2 analysis pipeline was judged as 'incomplete or missing' in 21 cases, and 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 re-analyst. For each issue, the panel member reviewed the evaluators' initial categorisation and their reasoning. Note that our aim in the project was to explore the sensitivity of analytical results to the analytical choices of the re-analysts. Hence, during the peer evaluation process, our goal was not to ensure that each analysis pipeline consisted of the ideal steps from every possible perspective, but to ensure that the steps of the analyses fell 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 categorised 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'.

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 '(1) 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 categorisation of results to be inadequate when they also judged that the conclusion does not follow from the results. It was often the case that verifying the legitimacy of the Task 1 conclusion also verified the legitimacy of the Task 1 categorisation.

Following the Task 2 analysis pipeline review, all initial ratings of '(1) Unacceptable' (n = 17) were revised to '(2) Acceptable but low quality'. 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, no changes were made to initial ratings following the review of code mismatches. 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, categorisation, or effect size.

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

***Resulting actions***

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

## Analysis methods

### *Marginal Effect Sizes*

The generalised Marginal Effects (gMEs) were calculated as specified by definition 3 of Kümpel and Hoffmann (2022). Specifically, the calculation of means was based on the empirical distribution of the data each analyst used to fit a given model or compute a test statistic. The standardised gME values were obtained by dividing each gME point estimate by the standard deviation of the target variable. To facilitate this calculation, it was necessary to replicate the analysis code of each analyst and extract the data after preprocessing, as well as draws from the posterior distribution or, alternatively, point estimates and variance-covariance matrices (for details see Kümpel & Hoffmann, 2022). Where applicable, precisely in a single instance, *t*-test analyses were redone by fitting a simple linear regression model, i.e., a linear regression with a single independent variable.

## Project timeline

The main milestones of the project were

- Start of the project Feb 10, 2021
- Recruitment of expert panel Feb 24, 2021
- Start of re-analyst recruitment Jan 21, 2022
- Start of re-analyses Nov 19, 2022
- Completion of the empirical work Oct 22, 2024