Results

## General descriptives

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.

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

Out of the submitted analyses 4 of them were omitted from the summary analysis their analysis as their analysis failed the peer evaluation.

As a result, we ended up with 505 re-analyses, submitted by 458 co-analysts.

Although we invited more than 5 co-analysts to each of the 100 studies, due to drop-outs and peer evaluation exclusions the final number of completed analyses ranged between 4 and 7. Table 1 shows the the distribution of analyses per studies.

| Number of Completed Analyses | Number of Studies |
| --- | --- |
| 4 | 14 |
| 5 | 69 |
| 6 | 15 |
| 7 | 2 |

## Basic demographics of the co-analysts

Out of all the co-analysts who submitted their work by the deadline, there were professors, associate professors, assistant professors, post-doctoral researchers, doctoral students, from other academic/research positions.

The gender distribution of the co-analysts is as follows: 117 female, 333 male, 1 other, and 7 didn’t want to respond to this question.

The age distribution of the co-analysts is depicted in [Figure 1](#fig-age-plot). 376 young adults (-39 years); 81 middle-aged adults (40-59 years); and no old adults (60- years).

|  |
| --- |
| Figure 1: The figure shows the distribution of the analysts’ age. When an analyst submitted more than one re-analysis with more than a year apart, we only kept their age at the time of their first submission. Moreover, one analyst was excluded because they did not disclose their age. |

Regarding the highest level of education, 1 reported High-school diploma or equivalent, 18 co-analysts had Bachelor’s degree or equivalent, 135 Master’s degree or equivalent, 304 had Doctoral degree or equivalent. In case the analysts completed more than one re-analysis and they advanced in their studies by the time of their second analysis, we only kept their first response for this comparison.

The country of residence of the co-analysts is shown on the map on [Figure 2](#fig-country-plot). Regarding the continents, 1 co-analyst was from Africa, 27 were from Asia, 15 from Oceania, 296 from Europe, 113 from North America, 6 from South America.

|  |
| --- |
| Figure 2: The figure shows the analysts’ country of residence. When an analyst submitted more than one re-analysis and they moved between the submissions, we only kept their first response. |

We asked the co-analysts which discipline is the closest to their research area. The following Table summarizes the distribution of their disciplinary orientation. Co-analysts from Psychology and Economics disciplines participated in the highest ratio in this study.

| Discipline | Count | Percentage |
| --- | --- | --- |
| Psychology | 265 | 57.86 |
| Economics | 74 | 16.16 |
| Political Science | 34 | 7.42 |
| Business Studies | 27 | 5.90 |
| Sociology | 19 | 4.15 |
| Computer Science/Statistics/Data Science | 16 | 3.49 |
| Public Policy | 3 | 0.66 |
| Anthropology | 1 | 0.22 |
| International Relations | 1 | 0.22 |
| Other | 18 | 3.93 |
| Note: Whenever the respondents provided more than one field we only kept their first responses. | | |

The distribution of the years of experience with data analysis is depicted on [Figure 3](#fig-experience-years-plot). The median time of experience with data analysis was 8 years among our co-analysts.

|  |
| --- |
| Figure 3: The figure shows the analysts’ years of experience with data analysis. When an analyst submitted more than one re-analysis and a year passed between the responses we only kept their first response. |

We asked our co-analysts how regularly they perform data analysis. [Figure 4](#fig-analysis-frequency) shows that the most frequent category was 2-3 times a week.

|  |
| --- |
| Figure 4: The figure shows how regularly the analysts perform data analysis. |

We also asked them how they rated their level of expertise in data analysis between Beginner (1) and Expert (10). The distribution on [Figure 5](#fig-self-rating-plot) shows that the most prevalent answer was 8 .

|  |
| --- |
| Figure 5: The figure shows the analysts’ self-rated level of expertise in data analysis. When an analyst submitted more than one re-analysis we only kept their first response. |

In 8.12 % (41 out of 505) of the cases, the co-analysts were familiar with the paper that the provided dataset belongs to before beginning their work on the project.

All co-analyst reported that they have not communicated about the details of their analysis with other co-analysts working with the same dataset.

We asked the co-analysts what programming language/software/tool they used in their data analysis during Task 1 and Task 2. The following figure indicates that R (62.49%), STATA (16.8%), and SPSS (6.99%) were the most popular responses. [Figure 6](#fig-software) shows the distribution of all the responses.

|  |
| --- |
| Figure 6: The figure shows which software the analysts used for their re-analysis tasks. In case an analyst completed multiple re-analyses or reported the use of multiple software we kept all their responses for this figure. The figure shows only software that was used by more than 1% of the analysts. |

## Descriptives of the statistical analyses

A difference in Task 2 compared to Task 1 was that the co-analysts received some constraints for their analysis to make their result comparable to a single result in the original study (see Methods for more details).

In Task 2, when we asked the co-analysts to present one main statistical result, in 97.62% of the analyses (493 out of 505), conclusion was based on the p-value. Bayes Factor was used in 2.38% of the cases (12 out of 505).

For 47.72 % (241 out of 505) of the analyses, the co-analysts reported having to make additional calculations in Task 2 compared to Task 1. In the remaining 52.28% (264 out of 505) of the cases, the co-analysts indicated that despite the requirements of the instructions, they could conduct the same analyses as in Task 1.

In Task 2, 12.67% of the results (64 out of 505) were in the opposite direction as claimed by the original study, disregarding whether the effect was conclusive/significant.

The co-analysts were asked to estimate the time they spent performing Task 1 and Task 2 together. The median value of their response is 6 hours ([Figure 7](#fig-total-hours)).

|  |
| --- |
| Figure 7: The figure shows the total hours the analyst spent on Task 1 and Task 2 together. In case an analyst completed multiple re-analyses, we kept all their responses for this figure. One response was excluded due to being an outlier (999 hours). |

## Peer evaluation

### Peer evaluators

#### Basic demographics of the peer evaluators

[Figure 8](#fig-evaluator-years) shows that most peer evaluators have many years of experience with conducting statistical analysis.

|  |
| --- |
| Figure 8: The figure shows the peer evaluators’ years of experience with data analysis. When a peer evaluator submitted more than one evaluation and a year passed between the responses, we kept only their first response. |

[Figure 9](#fig-evaluator-analysis-frequency) indicates that peer evaluators regularly perform data analysis.

|  |
| --- |
| Figure 9: The figure shows how regularly the peer evaluators perform data analysis. |

[Figure 10](#fig-evaluator-expertise) indicates that most peer evaluators rate themselves close to expert level in data analysis.

|  |
| --- |
| Figure 10: The figure shows the peer evaluators’ self-rated level of expertise in data analysis. When a peer evaluator submitted more than one re-analysis, we kept only their first response. |

### Peer evaluations

#### Descriptives of peer evaluations

In total, we received X peer evaluation reports. X peer evaluation was removed because the ID of the analyst was not provided, and as such, we could not verify with certainty which re-analysis was being evaluated leaving us with a total of X peer evaluation reports on Y different papers. After the panel member review of the peer evaluations (see Peer Evaluation: Review and Decisions’ supplement for all decisions and reasoning behind each case), the final result of the peer evaluation was the following.

task1\_pipeline\_acceptable In 99.8% of the cases (488 out of 489), the Task 1 analysis pipeline was categorized as acceptable.  
task1\_conclusion\_follows\_results In 99.59% of the cases (487 out of 489), the conclusions adequately followed from the results of the Task 1 analyses.

One of the Task 2 analysis pipelines was incomplete, the other 488 analyses were all categorized as acceptable.

In X% of the cases (X out of X), the conclusions adequately followed from the results of the Task 2 analyses.

HERE

99.39% (486 out of 489) the co-analyst’s self-categorization (meaning xxx) of the results was adequate.

X analytical reproducibility checks was successfully conducted which identified X mismatches.

### Old

For those analyses where there were more than one peer evaluations, for 49.51% (101 out of 204) of the analysis the evaluators disagreed on the analytical pipeline for task 1, and 53.43% (109 out of 204) for task 2.

99.8% (488 out of 489) of acceptable analysis pipelines (Task 1) - the outcome of the procedure

7 of peer elevators’ responses we need to adjust for this variable (comment from Marci: we only adjusted the unacceptable responses to acceptable but with low quality I think we can be explicit about that)

99.8% (488 out of 489) analysis pipelines (Task 2) were acceptable, and % ( out of ) analysis pipelines (Task 2) were incomplete

38 responses of peer evaluations we need to adjust for this variable (comment form marci: again we only modified the unacceptable to acceptable)

For those analyses where there were more than one peer evaluator, 0.49% (1 out of 204) of evaluators disagreed on the adequacy of the conclusions.

99.39% (486 out of 489) the co-analyst’s self-categorization of the results was adequate.

35 peer evaluations we need to adjust

99.59% (487 out of 489) conclusions adequately followed from the results of the analysis for Task 1.

% (x out of y) of cases where the correction of the self-categorization of the conclusion was necessary (comment from marci: this is 0 right now as we discussed, harry will go through them)

Nr. of analytical reproducibility checks:

74.3% (185 out of 249 of the analytical reproducibility checks were successful

## Inferential robustness: The robust of the conclusions to analytical choices published in social sciences

Do different analysts arrive at the same conclusions as the analysts of the original study?

### Task 1 Survey results

In Task 1, the co-analysts were asked to conduct any statistical analysis to arrive at a single conclusion. Out of 100 re-analysed studies, the conclusions of 34 (34%) remained robust to independent re-analysis, so all assigned co-analysts arrived at the same conclusion as reported in the article of the original study (inferential robustness; see [Figure 11](#fig-conclusions-main-robustness)).

|  |
| --- |
| Figure 11: The figure shows the proportion of the inferentially robust and not robust studies. |

**?@fig-conclusions-mainshows** the histogram display of the different and identical conclusions resulting from the re-analysis of each of the studies.

|  |
| --- |
| Figure 12: The figure shows the percentage of identical, inconclusive, and different conclusions for each study. Study numbers correspond to studies listed in Table S1. |

Across all the re-analyses, 73.47 % (371 out of 505) of them arrived at the same conclusion; 24.36% (123 out of 505) to no effects, and 2.18% (11 out of 505) to opposite effect compared to the original conclusion.

#### Inferential robustness by discipline

We were interested to see whether the above results show a different pattern when inspecting them in different disciplines. [Figure 13](#fig-discipline-robustness) shows that for the fields with more than 10 studies in our collection (Economics, Political Science, and Psychology) the pattern was comparably similar. We found no outstanding differences between the discipline regarding the percentage of different and identical conclusions either (see [Figure 14](#fig-conclusions-discipline)).

|  |
| --- |
| Figure 13: The figure shows the inferential robustness of the studies by major disciplines (more than 10 studies in our collection). |

|  |
| --- |
| Figure 14: The figure shows the percentage of identical, inconclusive, and different conclusions of the studies by major disciplines. The figure displays the count of re-analyses next to each discipline name. |

#### Inferential robustness by study type (observational, experimental)

Here, we were interested to see whether these results show a different pattern when separating them by study type. [Figure 15](#fig-studytype-robustness) illustrates that nearly half of the results from experimental studies remained robust upon independent re-analysis, whereas only one-third of observational studies yielded robust conclusions. Moreover, [Figure 16](#fig-conclusions-studytype) indicates that, for both study types, the majority of the re-analyses reached the same conclusions as the original study.

|  |
| --- |
| Figure 15: The figure shows the inferential robustness of the studies by study type (experimental or observational). The figure displays the count of re-analyses next to each study type name. |

|  |
| --- |
| Figure 16: The figure shows percentage of same conclusion, no effect/inconclusive, and opposite effect of the re-analyses by study type (experimental, observational). |

#### Inferential robustness by expertise (self-reported expertise in data analysis)

Here, we were interested to see whether these results show a different pattern when inspecting them along the reported expertise of the co-analysts. [Figure 17](#fig-conclusions-expertise) shows these results.

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
ℹ Please use `linewidth` instead.

|  |
| --- |
| Figure 17: The figure shows the percentage of same conclusion, no effect/inconclusive, and opposite effect of the re-analyses by self-rated expertise (on a scale of 1 (Beginner) to 10 (Expert)). The figure does not display the bottom two categories where fewer than 3 responses were collected for each. |

#### Inferential robustness by prior familiarity with the dataset

Here, we were interested to see whether these results show a different pattern when inspecting them along their prior familiarity with the dataset. [Figure 18](#fig-conclusions-familiarity) shows that for these results.

|  |
| --- |
| Figure 18: The figure shows the percentage of same conclusion, no effect/inconclusive, and opposite effect of the re-analyses by declared familiarity with the study. |

#### Inferential robustness by the level of confidence with the suitability of the analysis

|  |
| --- |
| Figure 19: The figure 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. |

#### Inferential robustness by the sample size

Here, we were interested to see whether these results show a different pattern when considering sample size. [Figure 20](#fig-conclusions-samplesize) shows that for that…,

|  |
| --- |
| Figure 20: This raincloud figure shows the distribution of the sample sizes of the re-analyses resulting in same conclusion, no effect/inconclusive, and opposite effects. |

#### Estimate robustness: robust of the statistical findings published in social sciences to analytical choices

A main question of our study was whether different analysts arrive at the same effect estimates (+/- 0.05 Cohen’s d) as the analyst of the original study? Figure 18 shows the distribution of the effect sizes of the original and next results. Figure 19 shows percentages of the effect sizes falling within the preset tolerance range (+/- 0.05 Cohen’s d) for each study.

|  |
| --- |
| Figure 21: The figure shows the effect size of the original result (black square) and the effect sizes of the re-analyses (green dot) for each study after conversions to Cohen’s d. For r missing\_original\_n studies, we could not determine the original effect size due to missing information. For r missing\_reanalysis\_n re-analyses, the reported effect size was not convertible to Cohen’s. The figure does not show r dplyr::filter(reanalysis\_data, effect\_size > 10 | effect\_size < -10) |> nrow() (r dplyr::pull(excluded\_es, excluded\_list)) re-analyzed effect sizes which are over 10 or smaller than -10 Cohen’s d. Study numbers correspond to studies listed in Table S1. |

|  |
| --- |
| Figure 22: The figure shows percentages of the effect sizes falling within the preset tolerance range (+/- 0.05 Cohen’s d) for each study. Study numbers correspond to studies listed in Table S1. |

We found that 96% (89 out of 93) of the studies contained at least one re-analysis result where the effect size was beyond the tolerance region (+/- 0.05 Cohen’s d) of the result of the original study. Out of the 392 available reanalysis effect sizes 67.86% (266) were outside of the tolerance region.

##### Estimate robustness by discipline

We were interested to see whether these robustness results show a different pattern when inspecting them in different disciplines. [Figure 23](#fig-effect-region-discipline) and [Figure 24](#fig-effect-robustness-discipline) show that for the major disciplines (>=10 studies).

|  |
| --- |
| Figure 23: The figure shows the percentage of re-analysis results falling within or outside of the tolerance region of the original results of the studies by major disciplines. The figure displays the count of re-analyses next to each discipline name. |

|  |
| --- |
| Figure 24: This raincloud figure shows for each major discipline the distribution of effect size estimate ranges (lowest to highest) calculated per study. |

##### Estimate robustness by study type (observational, experimental)

Here, we were interested to see whether these results show a different pattern when separating them by study type.

|  |
| --- |
| Figure 25: The figure shows the percentage of re-analysis results falling within or outside of the tolerance region of the original results of the studies by study type. The figure displays the count of re-analyses next to each discipline name. |

|  |
| --- |
| Figure 26: This raincloud figure shows for each study type the distribution of effect size estimate ranges (lowest to highest) calculated per study. |

##### Estimate robustness by expertise

Here, we were interested to see whether these results show a different pattern when inspecting them along the reported expertise of the co-analysts.

|  |
| --- |
| Figure 27: The figure shows the percentage of re-analysis results falling within or outside of the tolerance region of the original results of the studies by self-rated expertise (on a scale of 1 (Beginner) to 10 (Expert)). The figure displays the count of re-analyses next to each discipline name. |

##### Estimate robustness by prior familiarity with the dataset

Here, we were interested to see whether these results show a different pattern when inspecting them along their prior familiarity with the dataset. [Figure 28](#fig-effect-region-familiarity) shows that for these results.

|  |
| --- |
| Figure 28: The figure shows the percentage of re-analysis results falling within or outside of the tolerance region of the original results of the studies by declared familiarity with the study. |

##### Estimate robustness by the level of confidence with the suitability of the analysis

|  |
| --- |
| Figure 29: The figure shows |

##### Estiamte robustness by the sample size

Here, we were interested to see whether these results show a different pattern when considering sample size. The following [Figure 30](#fig-samplesize-region) shows no remarkable differences between the two categories.

|  |
| --- |
| Figure 30: The figure shows the distribution of sample sizes separately for re-analysis effect sizes falling within or outside of the tolerance region of the original results. In this figure, we did not include those studies where the original effect sizes were missing, and cases where the re-analysis effect size or sample size were missing. |

### Additional analyses

While Cohen’s d has the advantage of being easily interpretable and comparable across different analyses, it was designed to compare the means of two groups and its calculation relies on assumptions that can be compromised in more complex designs19. Recently, Kümpel & Hoffmann20 proposed a formal definition of generalized marginal effects (gMEs) that gives the calculation of effect size measures that are comparable across different statistical models. When standardized, the value of gMEs is equal to the value of Cohen’s d where the latter effect size measure is strictly applicable. Since we had not originally planned to calculate standardized gMEs, we did not collect all required analysis outputs to compute them. As a result, we calculated gMEs only for a sample of the 100 studies. Although this effect size calculation currently requires substantially more effort, it is recommended for future multi-analyst studies.

Hanna’s figure.