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 were omitted from the summary analysis 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 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 23 professors, 41 associate professors, 106 assistant professors, 107 post-doctoral researchers, 122 doctoral students, 59 from other academic/research positions, and from other 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 in order to make them linkable 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 p-value and 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 that they had to make additional calculations in the second task. 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 our 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)).

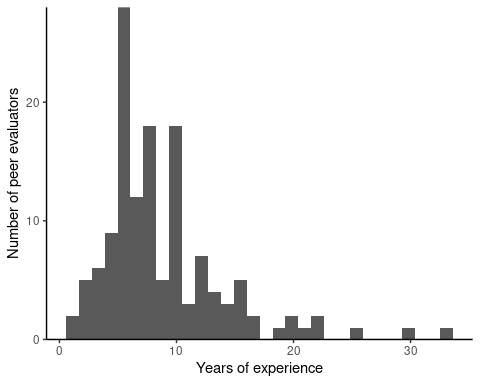
|  |
| --- |
| Figure 7: The figure shows the total hours the analyst spent on Task 1 and Task 2 together. In case an analyst completed multiple re-analyses, we kept all their responses for this figure. One response was excluded due to being an outlier (999 hours). |

## Peer evaluation

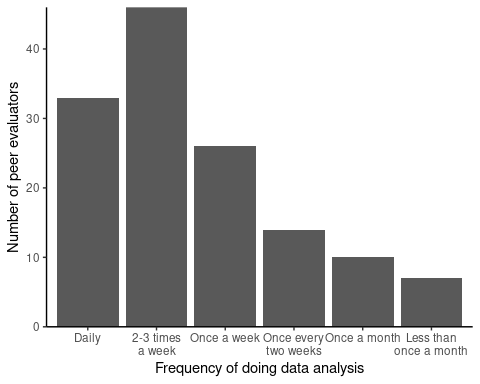
### Peer evaluators

Basic demographic info.

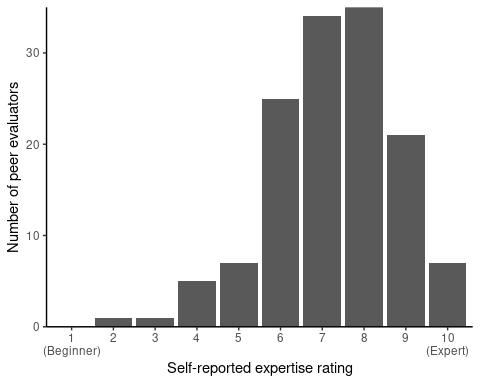
Experience with conducting statistical analysis:



Frequency of data analysis:



Self-reported expertise in data-analysis:



### Peer evaluations

Nr. of peer evaluations:

Descriptives of peer evaluations.

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

## How robust are 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 to a single conclusion. Out of 100 re-analysed studies, the conclusions of 34 (34%) remained robust to independent re-analysis, so that all assigned co-analysts arrived at the same conclusion as reported in the article of the original study.

|  |
| --- |
| Figure 8: The figure shows |

[Figure 9](#fig-conclusions-main) shows the histotrophic display of the different and identical conclusions resulting from the re-analysis of each of the studies.

|  |
| --- |
| Figure 9: The figure shows the histrophic display of the type of conclusions resulting from the re-analysis of each study. |

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

#### Inferential robustness by discipline

We were interested to see whether these results show a different pattern when inspecting them in different disciplines. [Figure 10](#fig-discipline-robustness) shows that for the major fields with more than 10 papers (Economics, Political Science, and Psychology) the pattern was comparably similar. We found no outstanding differences between the fields for the percentage of different and identical conclusions neither (see [Figure 11](#fig-conclusions-discipline)).

|  |
| --- |
| Figure 10: The figure shows the inferential robustness of the studies by major fields (more than 10 papers). |

|  |
| --- |
| Figure 11: The figure shows percentage of identical, inconclusive, and different conclusion of the studies by major fields. The figure displays the count of re-analyses next to each field 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 12](#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 13](#fig-conclusions-studytype) indicates that, for both study types, the majority of the re-analyses reached the same conclusions as the original study.

|  |
| --- |
| Figure 12: 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 field name. |

|  |
| --- |
| Figure 13: 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. The following figure shows these results.

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

|  |
| --- |
| Figure 14: The figure shows … The figure does not display the bottom two categories where for each less than 3 responses were collected. |

#### 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. The following figure shows that for these results.

|  |
| --- |
| Figure 15: The figure shows |

#### Inferential robustness by the suitability of their self-judged analysis

|  |
| --- |
| Figure 16: The figure shows |

#### Inferential 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 shows that for that…,

|  |
| --- |
| Figure 17: The figure shows |

#### How inferentially robust are 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?

The figure does not show 3 (study 042: 13.147; study 053: 60.578; study 057: 37.234) re-analyzed effect sizes which are over 10 or smaller than -10 Cohen’s d.

|  |
| --- |
| Figure 18: The figure shows |

|  |
| --- |
| Figure 19: The figure shows |

Here, we were interested in what percentage of the new effect sizes were beyond the tolerance region (+/- 0.05 Cohen’s d). 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 reanalyses effect sizes with available tolerance regions 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 fields. The following figure shows that for the major fields (>=10 studies).

|  |
| --- |
| Figure 20: The figure shows |

|  |
| --- |
| Figure 21: The figure shows |

##### 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 22: The figure shows |

|  |
| --- |
| Figure 23: The figure shows |

##### Estimate 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 24: The figure shows |

##### 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 25](#fig-effect-region-familiarity) shows that for these results.

|  |
| --- |
| Figure 25: The figure shows… |

##### Estimate robustness by the suitability of their self-judged analysis

|  |
| --- |
| Figure 26: 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 27](#fig-samplesize-region) shows that for that…,

|  |
| --- |
| Figure 27: The figure shows how the sample size influences whether the re-analysis effect sizes were within the tolerance region. For the figure we did not inlcude: those studies were the original effect sizes were missing, and cases where the re-analysis effect size or sample size were missing. |