General descriptives

## General

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 datasets 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Out of the submitted analyses, \_ were withdrawn, and 4 were omitted from the summary analysis their analysis failed the peer evaluation.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.

## Task 1 Survey results

Out of all the co-analysts who submitted their work by the deadline, there were 23 professors, 41 associate 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-ages 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, co-analysts had Bachelor’s degree or equivalent, Master’s degree or equivalent, had Doctoral degree or equivalent; and reported other degree. 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 continents, 1 co-analysts were from Africa, 27 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 summarises 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 |
| Other | 18 | 3.93 |
| Computer Science/Statistics/Data Science | 16 | 3.49 |
| Public Policy | 3 | 0.66 |
| Anthropology | 1 | 0.22 |
| International Relations | 1 | 0.22 |
| Note: Whenever the respondents provided more than one field we only kept their first responses. | | |

|  |
| --- |
| 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. |

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 4: The figure shows how regularly the analysts perform data analysis. When an analyst completed multiple re-analyses we kept all their responses for this figure. |

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

|  |
| --- |
| 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. |

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 .

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

No co-analysts reported that they 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. We only included software that was used by more than 1% of the analysts on the figure. |

## Task 2 Survey results

In Task 2, when we asked the co-analysts to present one main statistical result, 97.62% of them (493 our of 505) based their conclusion on p-value and 2.38% of them (12 our of 505) used Bayes Factor.

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. 47.72 % (241 out of 505) the co-analysts reported that they had to make additional calculations in the second task. In 52.28% (264 out of 505) the co-analysts indicated that despite the limitations in the instructions, they received the same result in Task 2 and 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.

|  |
| --- |
| Figure 7: The figure shows the total hours the analyst spent on Task 1 and Task 2. 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. |

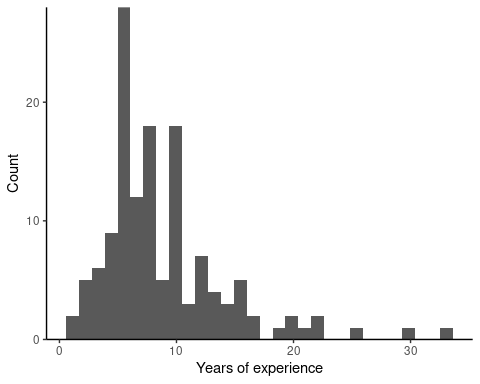
The co-analysts were asked to estimate the time they spent to perform Task 1 and Task 2 together. The median value of their response is 6 hours (Fig [Figure 7](#fig-total-hours)).

## Peer evaluation

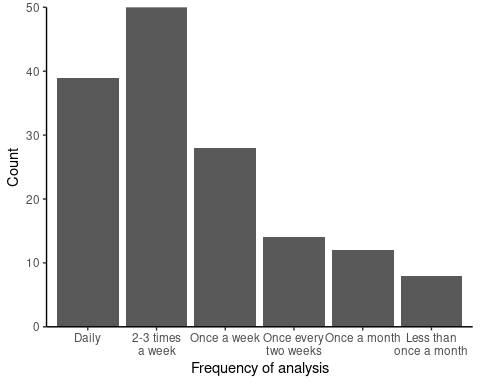
### Peer evaluators

Basic demographic info.

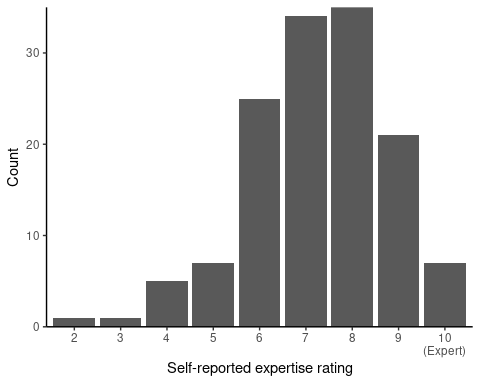
Experience with conducting statistical analysis:



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 49.02% (100 out of 204) of the analysis the evaluators disagreed on the analytical pipeline for task 1, and 57.84% (118 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)

95.5% (467 out of 489) analysis pipelines (Task 2) were acceptable, and 4.29% (21 out of 489) analysis pipelines (Task 2) were incomplete

17 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, 10.78% (22 out of 204) of evaluators disagreed on the adequacy of the conclusions.

92.23% (451 out of 489) the co-analyst’s self-categorization of the results was adequate.

0 peer evaluations we need to adjust

94.48% (462 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:

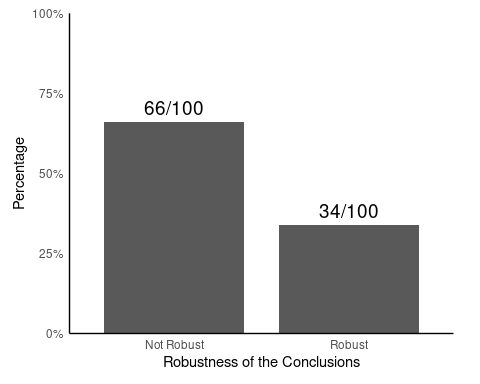
74.3% (185 out of 249 of the analytical reproducibility checks were successful

## How robust are conclusions published in social sciences to analytical choices?

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 2 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 the histrophic display of the conclusion resulting from the re-analysis of each study. |

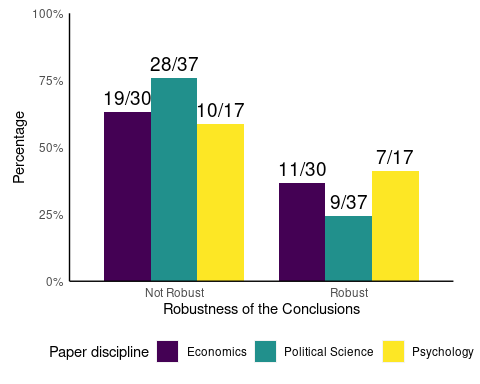
[Figure 8](#fig-conclusions-main) shows the histotrophic display of the different and identical conclusions resulting from the re-analysis of each of the studies.

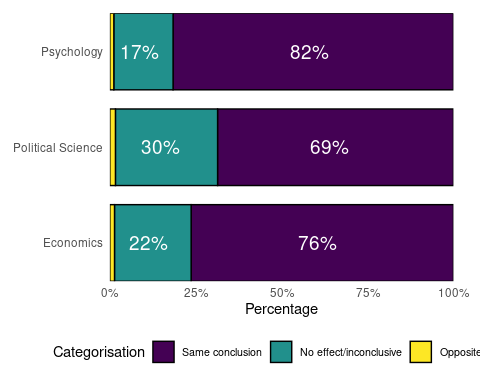
Across all the studies, 73.47 % (371 out of 505) of the re-analyses arrived to the same conclusion; 24.36% (123 out of 505) to no effects, and 2.18% (11 out of 505) to opposite effect compared to the original conclusion.

### Robustness

#### Robustness by field

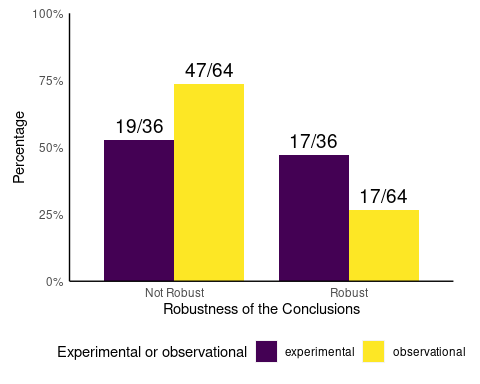
We were interested to see whether these results show a different pattern when inspecting them in different fields. The following figure shows that for the major fields (Psychology, Economics, and Political Science) the pattern were…,

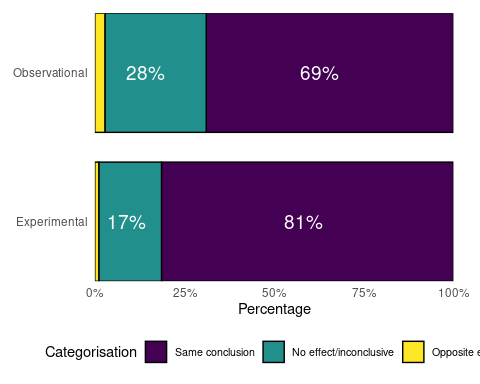




#### Robustness by study type (observational, experimental)

Here, we were interested to see whether these results show a different pattern when separating them by study type. The following figure shows that for that…,

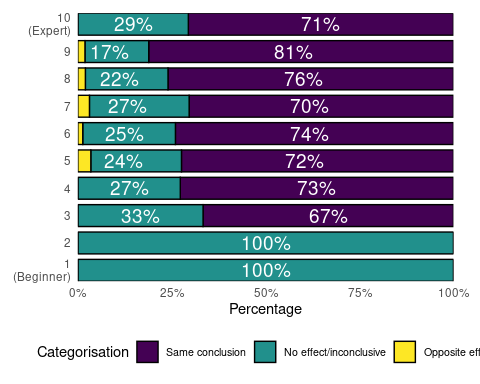




#### Robustness by expertise (self-reported expertise in data analysis)

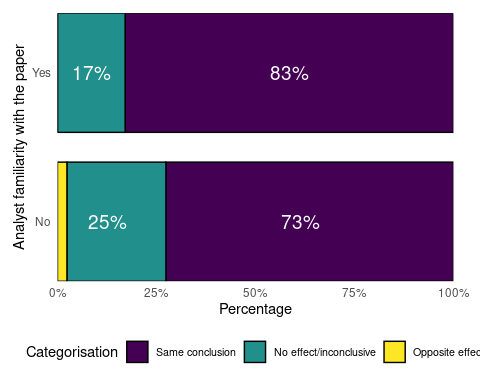
Here, we were interested to see whether these results show a different pattern when inspecting them along the reported expertise of the co-analysts. The following figure shows that for that…,

Saving 5 x 4 in image

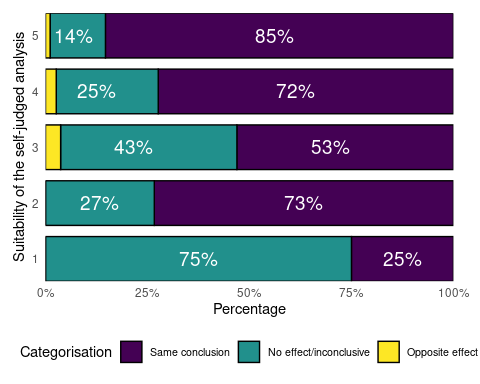


#### Robustness by prior familiarity with the dataset

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 that…,

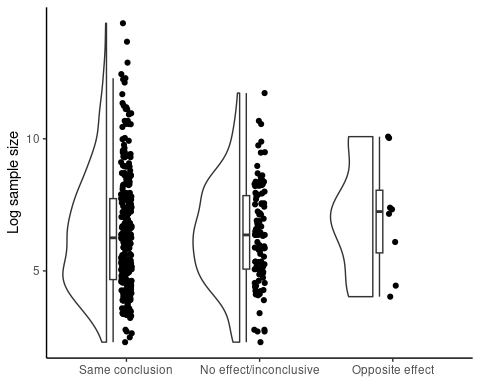


#### Robustness by the suitability of their self-judged analysis



#### 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…,



#### How 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?

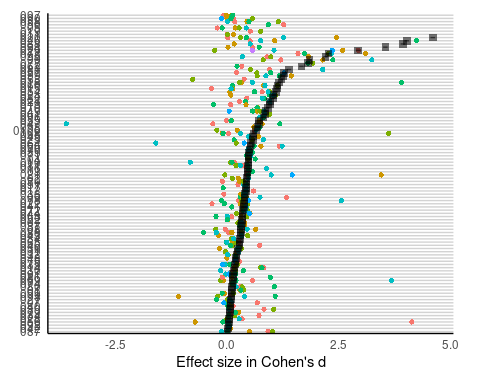
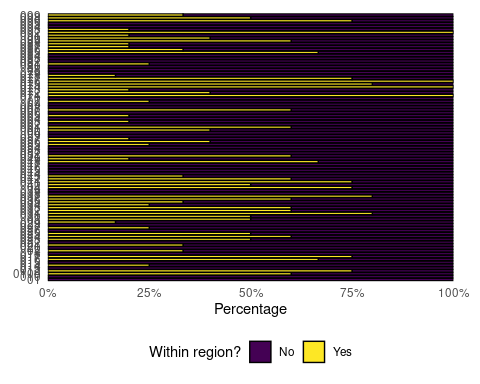


Fig X displays the calculated Cohen’s d-s for the re-analyses of each study and the Cohen’s d for the original study with a 0.05 tolerance region around it. For 7 papers the Cohen’s d for the original analysis cannot be calculated. In these cases, we could not calculate a tolerance region for the given paper, however, we show the available effect sizes for the re-analyses. For 89 analysis by the re-analyst the Cohen’s d cannot be calculated.

Fig X displays the calculated marginal ES-s for the re-analyses of each study (where they are available)

Fig X displays the calculated marginal ES-s AND corresponding Cohen’s d-s for the re-analyses of each study (where they are available)

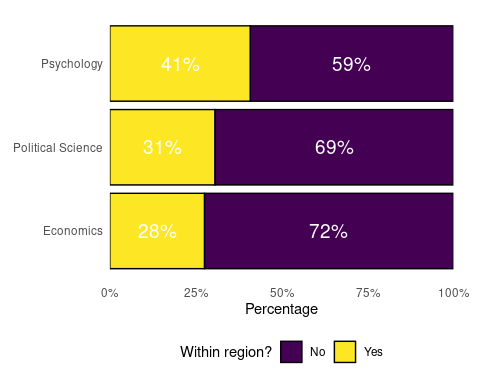


Here, we were interested what percentage of the new effect sizes were beyond the tolerance region (+/- 0.05 Cohen’s d). We found that % ( out of ) of the studies contained at least one re-analysis result where the effect size was beyond the tolerance region (+/- 0.05 Cohen’s d) of the result of the original study.

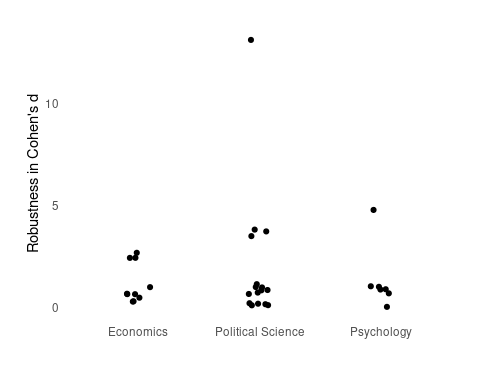
##### Robustness by field

We were interested to see whether these results show a different pattern when inspecting them in different fields. The following figure shows that for the major fields (>=10 studies) the pattern were…,

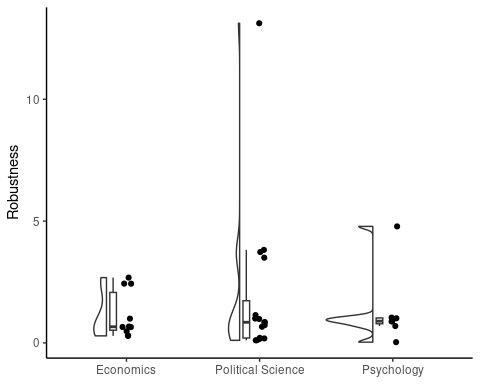
Saving 5 x 4 in image



`summarise()` has grouped output by 'paper\_id'. You can override using the  
`.groups` argument.  
Saving 5 x 4 in image

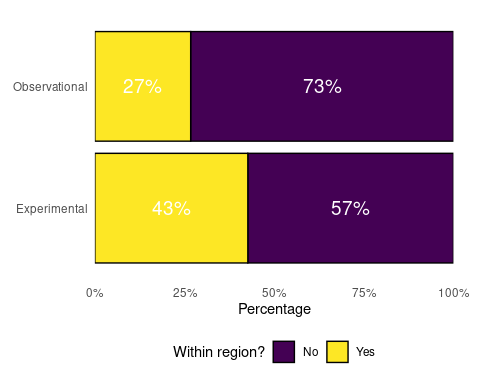


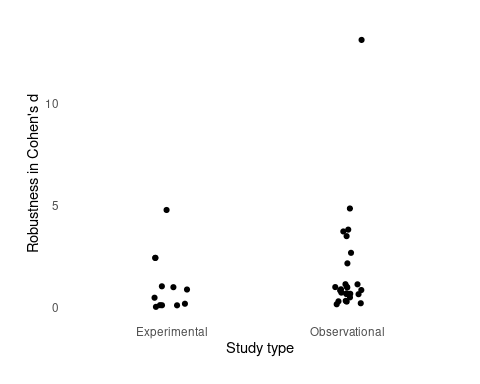
Saving 5 x 4 in image



##### Robustness by study type (observational, experimental)

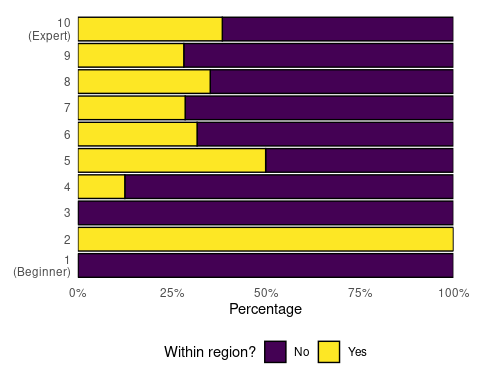
Here, we were interested to see whether these results show a different pattern when separating them by study type. The following figure shows that for that…,





##### 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 that for that…,

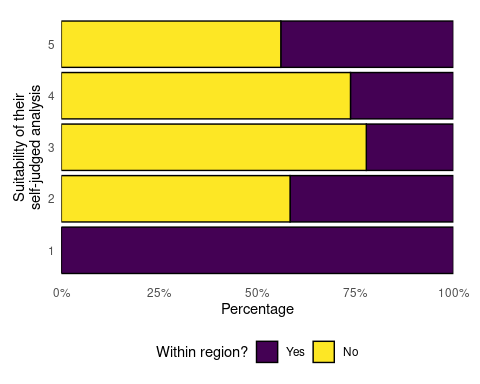


##### Robustness by prior familiarity with the dataset

Here, we were interested to see whether these results show a different pattern when inspecting them along their prior familiarity with the dataset. [Figure 9](#fig-effect-region-familiarity) shows that for that…,

|  |
| --- |
| Figure 9: The figure shows… |

##### Robustness by the suitability of their self-judged analysis



##### 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 10](#fig-samplesize-region) shows that for that…,

|  |
| --- |
| Figure 10: 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. |