|  |  |  |
| --- | --- | --- |
| Manuscript | | Multi100\_results code |
|  | |  |
| Only 31% of the independent re-analyses yielded the same result (within a tolerance region of +/- 0.05 Cohen’s *d*) as the original report | | We found that r dplyr::filter(within\_tolerance\_region, robust == "Inferentially not Robust") |> dplyr::pull(percentage)% (r dplyr::filter(within\_tolerance\_region, robust == "Inferentially not Robust") |> dplyr::pull(n) out of r dplyr::filter(within\_tolerance\_region, robust == "Inferentially not Robust") |> dplyr::pull(N)) 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 (@fig-effect-region-all). Out of the r dplyr::distinct(ind\_within\_tol\_reg, N) |> dplyr::pull(N) available reanalysis effect sizes r dplyr::filter(ind\_within\_tol\_reg, is\_within\_region == "Outside of tolerance region") |> dplyr::pull(percentage)% (r dplyr::filter(ind\_within\_tol\_reg, is\_within\_region == "Outside of tolerance region") |> dplyr::pull(n)) were outside of the tolerance region.  (NOT THE SAME) |
| Even with a four times broader tolerance region, this indicator did not go above 56% | | With a broader tolerance region (+/- 0.20 Cohen’s d) we found that from the available reanalysis effect sizes r dplyr::filter(ind\_within\_tol\_reg\_20, is\_within\_region == "Outside of tolerance region") |> dplyr::pull(percentage)% (r dplyr::filter(ind\_within\_tol\_reg\_20, is\_within\_region == "Outside of tolerance region") |> dplyr::pull(n)) were outside of the tolerance region.  (NOT THE SAME) |
| Regarding the conclusions drawn, only 34% of the studies remained analytically robust, meaning that all re-analysts reported evidence for the originally reported claim. | | In Task 1, the co-analysts were asked to conduct any statistical analysis to arrive at a single conclusion. Out of r distinct(conclusions\_main\_robustness\_data, N) |> pull(N) re-analysed studies, the conclusions of r filter(conclusions\_main\_robustness\_data, robust == "Inferentially robust") |> pull(percentage)% (r filter(conclusions\_main\_robustness\_data, robust == "Inferentially robust") |> pull(n)) 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 @fig-conclusions-main-robustness). |
| Using a more liberal definition of robustness produced comparable result (39% when >80% re-analysis agreement with the original conclusion defined analytical robustness). | | With alternative definitions of analytical robustness, this value was r filter(conclusions\_main\_robustness\_80\_data, robust == "Inferentially robust") |> pull(percentage)% when 80% and r filter(conclusions\_main\_robustness\_60\_data, robust == "Inferentially robust") |> pull(percentage)% when 60% reanalysis agreement with the original conclusion defined analytical robustness |
| After conducting 504 re-analyses with the involvement of 457 independent re-analysts on a selection of 100 social and behavioural studies, we conducted strictly explorative analyses on the results in order to describe the patterns in our findings. | | Throughout the project, r n\_analysis 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. |
| We found in 82% of the studies they reported different statistics regarding statistical test families (such as *t*-test, F-test, χ2) and their values (after rounding them to two decimal places). | | In Task 2 r dplyr::filter(proportion\_unique\_analysis, is\_unique) |> pull(percentage)% of the co-analyst used unique analytical pipelines based on the statistical test family and the value of the test statistics they arrived at. In total r pull(proportion\_unique\_paper, percentage\_all\_unique)% of the papers had completely unique reanalysis attempts. |
| Figure 2   * Effect size of the original analysis (black square; all represented as positive values) and the effect sizes of the re-analyses (green dot) for each study after conversions to Cohen’s *d*. The figure displays 447 re-analysis effect size estimates. For 57 re-analyses, the reported effect size was not convertible to Cohen’s *d*. Furthermore, the figure does not show 5 re-analysed effect sizes outside the [-5,5] range. For the seven studies listed at the bottom of the figure, we could not determine the original effect size due to missing information (for details see Supplementary information). Study numbers correspond to studies listed in <https://osf.io/mkwhn>. The studies are ordered by the size of the original effect size. * Proportion of effect sizes falling within the preset tolerance range (+/- 0.05 Cohen’s *d*) for each study. The studies are ordered by the displayed proportion. * 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 * Distributions of effect size estimate ranges (lowest to highest) calculated per study for each major discipline * Proportion 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. * Distribution of effect size estimate ranges (lowest to highest) calculated per study for observational and experimental studies * 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 from 1 (Beginner) to 10 (Expert)) * 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. * 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 could not include those studies where the original effect sizes were missing, and cases where the re-analysis effect size or sample size were missing | | IN CODE |
| A yellow and purple chart  Description automatically generated | | |
| We found that 96% of the studies for which we could obtain the original effect size (*n* = 93) 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 | We found that r dplyr::filter(within\_tolerance\_region, robust == "Inferentially not Robust") |> dplyr::pull(percentage)% (r dplyr::filter(within\_tolerance\_region, robust == "Inferentially not Robust") |> dplyr::pull(n) out of r dplyr::filter(within\_tolerance\_region, robust == "Inferentially not Robust") |> dplyr::pull(N)) 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 (@fig-effect-region-all). | |
| Out of the 417 available re-analysis effect sizes, 31.41% were inside the tolerance region | | Out of the r dplyr::distinct(ind\_within\_tol\_reg, N) |> dplyr::pull(N) available reanalysis effect sizes r dplyr::filter(ind\_within\_tol\_reg, is\_within\_region == "Outside of tolerance region") |> dplyr::pull(percentage)% (r dplyr::filter(ind\_within\_tol\_reg, is\_within\_region == "Outside of tolerance region") |> dplyr::pull(n)) were outside of the tolerance region. |
| We learned that even with a four times broader tolerance region (+/- 0.20 Cohen’s *d*), we would find that around 80% of the studies contained at least one re-analysis result where the effect size was beyond the tolerance region. | | With a broader tolerance region (+/- 0.20 Cohen’s d) we found that from the available reanalysis effect sizes r dplyr::filter(ind\_within\_tol\_reg\_20, is\_within\_region == "Outside of tolerance region") |> dplyr::pull(percentage)% (r dplyr::filter(ind\_within\_tol\_reg\_20, is\_within\_region == "Outside of tolerance region") |> dplyr::pull(n)) were outside of the tolerance region. |
| Further, out of the available 185 (after ….) reanalysis effect sizes (185) 55.64% were inside of this region | | With a broader tolerance region (+/- 0.20 Cohen’s d) we found that from the available reanalysis effect sizes r dplyr::filter(ind\_within\_tol\_reg\_20, is\_within\_region == "Outside of tolerance region") |> dplyr::pull(percentage)% (r dplyr::filter(ind\_within\_tol\_reg\_20, is\_within\_region == "Outside of tolerance region") |> dplyr::pull(n)) were outside of the tolerance region.  (NOT THE SAME) |
| Re-analysts reported that they were familiar with the original study in less than 10% of cases | | In r filter(familiar\_with\_paper\_data, familiar\_with\_paper == "Yes") |> pull(percentage) % (r filter(familiar\_with\_paper\_data, familiar\_with\_paper == "Yes") |> pull(n) out of r filter(familiar\_with\_paper\_data, familiar\_with\_paper == "Yes") |> pull(N)) of the cases, the co-analysts were familiar with the paper that the provided dataset belongs to before beginning their work on the project. |
| Moreover, there was no more than 5% difference in robustness between those who were and those who were not familiar with the original study | | @fig-conclusions-familiarity |
| The mean effect size of the original results is 0.72 (Median = 0.43), whereas for the re-analysis it is 0.48 (Median = 0.34) Cohen’s *d* computed on Cohen’s *d*s =< 5. | | Chunk 67 (probably) |
| Figure 3   * The thin diagonal line represents an ideal case when the re-analysis effect sizes are equal to original effect size, the thick line shows the best-fitting (least squares) line of the displayed dots. Density plots of original and re-analysis effect sizes are parallel to their respective axis. * Figure **a** show effect sizes Cohen’s *d* =< 5 * Figure **b** show effect sizes Cohen’s *d* =< 1 | | Chunk 67 |
|  | | |
| Out of 100 re-analysed claims, 34% were robust to independent re-analysis, such that all re-analysts reported that they found evidence for the originally reported claim. | | In Task 1, the co-analysts were asked to conduct any statistical analysis to arrive at a single conclusion. Out of r distinct(conclusions\_main\_robustness\_data, N) |> pull(N) re-analysed studies, the conclusions of r filter(conclusions\_main\_robustness\_data, robust == "Inferentially robust") |> pull(percentage)% (r filter(conclusions\_main\_robustness\_data, robust == "Inferentially robust") |> pull(n)) 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 @fig-conclusions-main-robustness). |
| With a more liberal definition of analytical robustness, this value was 39% when >80% re-analysis agreement with the original conclusion defined analytical robustness. | | With alternative definitions of analytical robustness, this value was r filter(conclusions\_main\_robustness\_80\_data, robust == "Inferentially robust") |> pull(percentage)% when 80% and r filter(conclusions\_main\_robustness\_60\_data, robust == "Inferentially robust") |> pull(percentage)% when 60% reanalysis agreement with the original conclusion defined analytical robustness |
| Across all individual re-analyses (*n* = 504), 73.61% of analyses were reported to arrive at the same conclusion as in the original investigation, 24.21% to no effects/inconclusive result, and 2.18% to the opposite effect as in the original investigation | | Across all the re-analyses, r filter(conclusions\_analysis\_data, categorisation == "Same conclusion") |> pull(percentage) % (r filter(conclusions\_analysis\_data, categorisation == "Same conclusion") |> pull(n) out of r filter(conclusions\_analysis\_data, categorisation == "Same conclusion") |> pull(N)) of them arrived at the same conclusion; r filter(conclusions\_analysis\_data, categorisation == "No effect/inconclusive") |> pull(percentage)% (r filter(conclusions\_analysis\_data, categorisation == "No effect/inconclusive") |> pull(n) out of r filter(conclusions\_analysis\_data, categorisation == "No effect/inconclusive") |> pull(N)) to no effects, and r filter(conclusions\_analysis\_data, categorisation == "Opposite effect") |> pull(percentage)% (r filter(conclusions\_analysis\_data, categorisation == "Opposite effect") |> pull(n) out of r filter(conclusions\_analysis\_data, categorisation == "Opposite effect") |> pull(N)) to opposite effect compared to the original conclusion. |
| Figure 4   * Proportion of same conclusion, no effect/inconclusive results, and opposite direction conclusions for each study * Proportion of inferentially robust results (i.e., all re-analyses arrived at the same conclusion for the given study) by major disciplines (more than 10 studies in our collection: Economics, Political Science, and Psychology) * Proportion of same effect, no effect/inconclusive results, and conclusions in the opposite direction of the original studies by major discipline. The number of re-analyses is displayed below each discipline * Proportion of inferentially robust results by study design (experimental vs. observational). The number of re-analyses is given below each study design. **e,** Proportion of same conclusion, no effect/inconclusive, and opposite effect of the re-analyses by study type (experimental, observational). * Proportion 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)). * Proportion of inferentially robust studies by the acceptability of the analysis pipelines according to the peer evaluators. For this figure, we included only studies with more than one peer evaluation and where the peer evaluators agreed on their rating. The figure shows only the rating options with 5 or more re-analyses in that category. * Proportion of same conclusion, no effect/inconclusive, and opposite effect of the re-analyses by declared familiarity with the study. * Distribution of the sample size of the re-analyses resulting in the same conclusion, no effect/inconclusive, and opposite effects. Sample size values were available for 345 re-analyses. | | IN CODE |
| A group of different colored bars  Description automatically generated with medium confidence | | |
| Although those familiar with the original study did report the same conclusion in a higher proportion than those who were not familiar, 17% of their re-analyses still indicated a conclusion different from the original one | | @fig-conclusions-familiarity |
| Overall, when independent researchers analysed the same research question on the original data, 34% of studies remained robust to independent re-analysis in the sense that all re-analysts arrived at the same conclusion as the original analyst or analyst team. | | In Task 1, the co-analysts were asked to conduct any statistical analysis to arrive at a single conclusion. Out of r distinct(conclusions\_main\_robustness\_data, N) |> pull(N) re-analysed studies, the conclusions of r filter(conclusions\_main\_robustness\_data, robust == "Inferentially robust") |> pull(percentage)% (r filter(conclusions\_main\_robustness\_data, robust == "Inferentially robust") |> pull(n)) 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 @fig-conclusions-main-robustness). |
| Importantly, the difference in statistical results altered the inferential conclusions in 26% of the re-analyses | | In Task 2, r filter(direction\_of\_result\_data, direction\_of\_result == "Opposite as claimed by the original study") |> pull(percentage)% of the results (r filter(direction\_of\_result\_data, direction\_of\_result == "Opposite as claimed by the original study") |> pull(n) out of r filter(direction\_of\_result\_data, direction\_of\_result == "Opposite as claimed by the original study") |> pull(N)) were in the opposite direction as claimed by the original study, disregarding whether the effect was conclusive/significant. |
| Analytic reproducibility was tested in cases when both original data and code were available (n = 63), or when the original data were available but the original code had to be adapted by the SCORE team in order to successfully reproduce the result (n = 7). If data were available but the original code was not, SCORE sourced a collaborating lab to generate new analytic code for the reproduction (n = 10). If data and code were not available, the collaborating lab used the secondary source data which were shared upon request (acquired by SCORE), alongside newly generated analytical code for the reproduction (n = 11). Some reproductions were never attempted (n = 9) | | NOT IN CODE |
| In practice, for logistical reasons, this rule was applied in all but seven cases (i.e., 98.6% of peer evaluations were carried out on a dataset that was different to the dataset they analysed themselves). | | # Check if there are peer evaluators who evaluated an analysis for the same paper they analyzed  peer\_eval |>  dplyr::select(evaluator\_id, paper\_id) |>  rename(evaluator\_paper\_id = paper\_id) |>  left\_join(distinct(processed, paper\_id, analyst\_id), by = c("evaluator\_id" = "analyst\_id"), relationship = "many-to-many") |>  mutate(match = case\_when(  paper\_id == evaluator\_paper\_id ~ TRUE,  paper\_id != evaluator\_paper\_id ~ FALSE  )) |>  filter(match) |
| 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. | | As a response to our recruitment call, r nrow(analyst\_signed\_up) researchers signed up to participate in our study. Out of these volunteers, r analyst\_submitted signed up to analyse at least one dataset and submitted their work by the deadline or an extended deadline. |
| After conducting 504 re-analyses with the involvement of 457 independent re-analysts on a selection of 100 social and behavioural studies, we conducted strictly explorative analyses on the results in order to describe the patterns in our findings. (NOT THE SAME) | | Throughout the project, r n\_analysis 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. |
| NOT IN MANUSCRIPT | | Out of the submitted analyses r nrow(dplyr::filter(processed, !peer\_eval\_pass)) of them was omitted from the summary analysis as its analysis failed the peer evaluation and an additional r nrow(dplyr::filter(processed, !incomplete\_response\_pass)) analyses were excluded due to incomplete responses. |
| NOT IN MANUSCRIPT | | As a result, we ended up with r nrow(processed) re-analyses, submitted by r final\_n\_analyst co-analysts. |
| NOT IN MANUSCRIPT | | 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 r min(dplyr::count(processed, paper\_id)$n) and r max(dplyr::count(processed, paper\_id)$n). Table 1 shows the distribution of of the number of analyses for individual studies. |
| NOT IN MANUSCRIPT | | **Table 1. The Distribution of the Number of Analyses for Studies** |
| NOT IN MANUSCRIPT | | Out of all the co-analysts who submitted their work by the deadline, there were r dplyr::filter(position, position == "Professor") |> dplyr::pull(n) professors, r dplyr::filter(position, position == "Associate Professor") |> dplyr::pull(n) associate professors, r filter(position, position == "Assistant Professor") |> dplyr::pull(n) assistant professors, r dplyr::filter(position, position == "Post-Doc Researcher") |> dplyr::pull(n) post-doctoral researchers, r dplyr::filter(position, position == "Doctoral Student") |> dplyr::pull(n) doctoral students, r dplyr::filter(position, position == "Other academic/research position") |> dplyr::pull(n) from other academic/research positions. |
| NOT IN MANUSCRIPT | | The gender distribution of the co-analysts is as follows: r dplyr::filter(gender, gender == "Female") |> dplyr::pull(n) female, r dplyr::filter(gender, gender == "Male") |> dplyr::pull(n) male, r dplyr::filter(gender, gender == "Non-binary") |> dplyr::pull(n) other, and r dplyr::filter(gender, gender == "Prefer not to say") |> dplyr::pull(n) didn't want to respond to this question. |
| NOT IN MANUSCRIPT | | The age distribution of the co-analysts is depicted in @fig-age-plot. r dplyr::filter(age\_group, age\_group == "young") |> dplyr::pull(n) young adults (-39 years); r dplyr::filter(age\_group, age\_group == "middle") |> dplyr::pull(n) middle-aged adults (40-59 years); and no old adults (60- years). |
| NOT IN MANUSCRIPT | | Regarding the highest level of education, r dplyr::filter(education, education == "High-school diploma or equivalent") |> dplyr::pull(n) co-analyst reported High-school diploma or equivalent, r dplyr::filter(education, education == "Bachelor's degree or equivalent") |> dplyr::pull(n) co-analysts had Bachelor's degree or equivalent, r dplyr::filter(education, education == "Master's degree or equivalent") |> dplyr::pull(n) Master's degree or equivalent, r dplyr::filter(education, education == "Doctoral degree or equivalent") |> dplyr::pull(n) 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 kept only their first response for this comparison. |
| NOT IN MANUSCRIPT | | The country of residence of the co-analysts is shown on the map on @fig-country-plot. Regarding the continents, r dplyr::filter(continent, continent == "Africa") |> dplyr::pull(N) co-analyst was from Africa, r dplyr::filter(continent, continent == "Asia") |> dplyr::pull(N) were from Asia, r dplyr::filter(continent, continent == "Oceania") |> dplyr::pull(N) from Oceania, r dplyr::filter(continent, continent == "Europe") |> dplyr::pull(N) from Europe, r dplyr::filter(region, UNREGION1 == "Northern America") |> dplyr::pull(N) from North America, r dplyr::filter(region, UNREGION1 %in% c("Central America", "South America")) |> dplyr::summarise(sum(N)) |> dplyr::pull("sum(N)") from South America. |
| NOT IN MANUSCRIPT | | We asked the co-analysts which discipline is the closest to their research area. The following Table 2 summarizes the distribution of their disciplinary orientation. Co-analysts from r dplyr::slice(analyst\_discipline, 1) |> dplyr::pull(discipline) and r dplyr::slice(analyst\_discipline, 2) |> dplyr::pull(discipline) disciplines participated in the highest ratio in this study. |
| NOT IN MANUSCRIPT | | **Table 2. The Distribution of Co-analysts’ Disciplinary Orientation** |
| NOT IN MANUSCRIPT | | The distribution of the years of experience with data analysis is depicted on @fig-experience-years-plot. The median time of experience with data analysis was r median(analyst\_experience\_years\_data$ years\_of\_experience) years among our co-analysts. |
| NOT IN MANUSCRIPT | | We asked our co-analysts how regularly they perform data analysis. @fig-analysis-frequency shows that the most frequent category was r dplyr::filter(analysis\_frequency\_count, n == max(n)) |> dplyr::pull(analysis\_frequency). |
| NOT IN MANUSCRIPT | | We also asked them how they rated their level of expertise in data analysis between Beginner (1) and Expert (10). The distribution on @fig-self-rating-plot shows that the most prevalent answer was r dplyr::filter(expertise\_self\_rating\_data, n == max(n)) |> dplyr::pull(expertise\_self\_rating) . |
| Re-analysts reported that they were familiar with the original study in less than 10% of cases  (NOT THE SAME) | | In r filter(familiar\_with\_paper\_data, familiar\_with\_paper == "Yes") |> pull(percentage) % (r filter(familiar\_with\_paper\_data, familiar\_with\_paper == "Yes") |> pull(n) out of r filter(familiar\_with\_paper\_data, familiar\_with\_paper == "Yes") |> pull(N)) of the cases, the co-analysts were familiar with the paper that the provided dataset belongs to before beginning their work on the project. |
| NOT IN MANUSCRIPT | | 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 slice(software\_data, 1) |> pull(software) (r slice(software\_data, 1) |> pull(percentage)%), r slice(software\_data, 2) |> pull(software) (r slice(software\_data, 2) |> pull(percentage)%), and r slice(software\_data, 3) |> pull(software) (r slice(software\_data, 3) |> pull(percentage)%) were the most popular responses. @fig-software shows the distribution of these responses. |
| NOT IN MANUSCRIPT | | In Task 2, when we asked the co-analysts to present one main statistical result, in r filter(p\_value\_or\_bayes\_data, p\_value\_or\_bayes == "p-value") |> pull(percentage)% of the analyses (r filter(p\_value\_or\_bayes\_data, p\_value\_or\_bayes == "p-value") |> pull(n) out of r filter(p\_value\_or\_bayes\_data, p\_value\_or\_bayes == "p-value") |> pull(N)), conclusion was based on the p-value. Bayes Factor was used in r filter(p\_value\_or\_bayes\_data, p\_value\_or\_bayes == "Bayes factor") |> pull(percentage)% of the cases (r filter(p\_value\_or\_bayes\_data, p\_value\_or\_bayes == "Bayes factor") |> pull(n) out of r filter(p\_value\_or\_bayes\_data, p\_value\_or\_bayes == "Bayes factor") |> pull(N)). |
| NOT IN MANUSCRIPT | | For r filter(additional\_calculations\_data, additional\_calculations == "Yes") |> pull(percentage) % (r filter(additional\_calculations\_data, additional\_calculations == "Yes") |> pull(n) out of r filter(additional\_calculations\_data, additional\_calculations == "Yes") |> pull(N)) of the analyses, the co-analysts reported having to make additional calculations in Task 2 compared to Task 1. In the remaining r filter(additional\_calculations\_data, additional\_calculations == "No, I already had the neccessary calculations in Task 1") |> pull(percentage)% (r filter(additional\_calculations\_data, additional\_calculations == "No, I already had the neccessary calculations in Task 1") |> pull(n) out of r filter(additional\_calculations\_data, additional\_calculations == "No, I already had the neccessary calculations in Task 1") |> pull(N)) of the cases, the co-analysts indicated that despite the requirements of the instructions, they could conduct the same analyses as in Task 1. |
| NOT IN MANUSCRIPT | | In Task 2, r filter(direction\_of\_result\_data, direction\_of\_result == "Opposite as claimed by the original study") |> pull(percentage)% of the results (r filter(direction\_of\_result\_data, direction\_of\_result == "Opposite as claimed by the original study") |> pull(n) out of r filter(direction\_of\_result\_data, direction\_of\_result == "Opposite as claimed by the original study") |> pull(N)) were in the opposite direction as claimed by the original study, disregarding whether the effect was conclusive/significant. |
| NOT IN MANUSCRIPT | | 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 r median(total\_hours\_data$total\_hours) hours (@fig-total-hours). |
| We found in 82% of the studies they reported different statistics regarding statistical test families (such as *t*-test, F-test, χ2) and their values (after rounding them to two decimal places). (NOT THE SAME) | | In Task 2 r dplyr::filter(proportion\_unique\_analysis, is\_unique) |> pull(percentage)% of the co-analyst used unique analytical pipelines based on the statistical test family and the value of the test statistics they arrived at. In total r pull(proportion\_unique\_paper, percentage\_all\_unique)% of the papers had completely unique reanalysis attempts. |
| NOT IN MANUSCRIPT | | @fig-evaluator-years shows that most peer evaluators have many years of experience with conducting statistical analysis. |
| NOT IN MANUSCRIPT | | @fig-evaluator-analysis-frequency indicates that peer evaluators regularly perform data analysis. |
| NOT IN MANUSCRIPT | | @fig-evaluator-expertise indicates that most peer evaluators rate themselves close to expert level in data analysis. |
| NOT IN MANUSCRIPT | | In total, we received r nrow(peer\_eval\_not\_reviewed) + 1 peer evaluation reports. One 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 r nrow(peer\_eval\_not\_reviewed) peer evaluation reports on r nrow(distinct(peer\_eval\_not\_reviewed, paper\_id)) 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. |
| NOT IN MANUSCRIPT | | At the end of the peer evaluation process, one analysis was deemed to contain an unacceptable analysis pipeline. 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 accurately follow on from the results, and the analysts self-categorization of the results were considered adequate. |
| NOT IN MANUSCRIPT | | r reproducibility\_checks\_n analytical reproducibility checks were successfully conducted which identified mismatches in r filter(reproducibility\_checks\_data, any\_code\_mismatches == "(4) I executed it and I found mismatches") |> pull(n) analyses. In all of these cases we verified that that the mismatches did not have a meaningful impact on the reported conclusion, categorization, or effect size. |
| Multiple places | | In Task 1, the co-analysts were asked to conduct any statistical analysis to arrive at a single conclusion. Out of r distinct(conclusions\_main\_robustness\_data, N) |> pull(N) re-analysed studies, the conclusions of r filter(conclusions\_main\_robustness\_data, robust == "Inferentially robust") |> pull(percentage)% (r filter(conclusions\_main\_robustness\_data, robust == "Inferentially robust") |> pull(n)) 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 @fig-conclusions-main-robustness). With alternative definitions of analytical robustness, this value was r filter(conclusions\_main\_robustness\_80\_data, robust == "Inferentially robust") |> pull(percentage)% when 80% and r filter(conclusions\_main\_robustness\_60\_data, robust == "Inferentially robust") |> pull(percentage)% when 60% reanalysis agreement with the original conclusion defined analytical robustness. For one study (1%), all co-analysts reported no effect or an inconclusive conclusion. |
| Figure 4a | | @fig-conclusions-main shows the histogram display of the different and identical conclusions resulting from the re-analysis of each of the studies. |
| Across all individual re-analyses (*n* = 504), 73.61% of analyses were reported to arrive at the same conclusion as in the original investigation; 24.21% to no effects/inconclusive result, and 2.18% to the opposite effect as in the original investigation | | Across all the re-analyses, r filter(conclusions\_analysis\_data, categorisation == "Same conclusion") |> pull(percentage) % (r filter(conclusions\_analysis\_data, categorisation == "Same conclusion") |> pull(n) out of r filter(conclusions\_analysis\_data, categorisation == "Same conclusion") |> pull(N)) of them arrived at the same conclusion; r filter(conclusions\_analysis\_data, categorisation == "No effect/inconclusive") |> pull(percentage)% (r filter(conclusions\_analysis\_data, categorisation == "No effect/inconclusive") |> pull(n) out of r filter(conclusions\_analysis\_data, categorisation == "No effect/inconclusive") |> pull(N)) to no effects, and r filter(conclusions\_analysis\_data, categorisation == "Opposite effect") |> pull(percentage)% (r filter(conclusions\_analysis\_data, categorisation == "Opposite effect") |> pull(n) out of r filter(conclusions\_analysis\_data, categorisation == "Opposite effect") |> pull(N)) to opposite effect compared to the original conclusion. |
| Figure 4b + Figure 4c | | We were interested to see whether the above results show a different pattern when inspecting them in different disciplines. @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 @fig-conclusions-discipline). |
| Figure 4d + Figure 4e | | Here, we were interested to see whether these results show a different pattern when separating them by study type. @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, @fig-conclusions-studytype indicates that, for both study types, the majority of the re-analyses reached the same conclusions as the original study. |
| Figure 4f | | Here, we were interested to see whether these results show a different pattern when inspecting them along the reported expertise of the co-analysts. @fig-conclusions-expertise shows these results. |
| Figure 4g | | @fig-subset-task1-pipeline shows the inferential robustness of the studies by the acceptability of the analysis pipelines according to the peer evaluators. |
| Figure 4h | | Here, we were interested to see whether these results show a different pattern when inspecting them along their prior familiarity with the dataset. @fig-conclusions-familiarity shows that for these results. |
| NOT IN MANUSCRIPT | | The following Table 3 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. |
| Figure 4i | | Here, in @fig-conclusions-samplesize we were interested to see whether these results show a different pattern when considering sample size. |
| Figure 2a | | 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-effect-main shows percentages of the effect sizes falling within the preset tolerance range (+/- 0.05 Cohen’s d) for each study. The figure displays r nrow(dplyr::filter(reanalysis\_data, !is.na(effect\_size))) re-analysis effect size estimates. For r missing\_reanalysis\_n re-analyses, the reported effect size was not convertible to Cohen’s d. The figure does not show r dplyr::filter(reanalysis\_data, effect\_size > 5 | effect\_size < -5) |> nrow() re-analysed effect sizes which are higher than 5 or lower than -5 Cohen's d (r dplyr::pull(excluded\_es, excluded\_list)). For the r missing\_original\_n studies listed in the bottom of the graph, we could not determine the original effect size due to missing information. |
| We found that 96% of the studies for which we could obtain the original effect size (*n* = 93) 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 (Fig. 2a). Out of the 417 available re-analysis effect sizes, 31.41% were inside the tolerance region. As a robustness test of our analysis, we explored the degree to which we would observe different results with different tolerance regions.  + Figure 2b | | We found that r dplyr::filter(within\_tolerance\_region, robust == "Inferentially not Robust") |> dplyr::pull(percentage)% (r dplyr::filter(within\_tolerance\_region, robust == "Inferentially not Robust") |> dplyr::pull(n) out of r dplyr::filter(within\_tolerance\_region, robust == "Inferentially not Robust") |> dplyr::pull(N)) 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 (@fig-effect-region-all). Out of the r dplyr::distinct(ind\_within\_tol\_reg, N) |> dplyr::pull(N) available reanalysis effect sizes r dplyr::filter(ind\_within\_tol\_reg, is\_within\_region == "Outside of tolerance region") |> dplyr::pull(percentage)% (r dplyr::filter(ind\_within\_tol\_reg, is\_within\_region == "Outside of tolerance region") |> dplyr::pull(n)) were outside of the tolerance region. |
| We learned that even with a four times broader tolerance region (+/- 0.20 Cohen’s *d*), we would find that around 80% of the studies contained at least one re-analysis result where the effect size was beyond the tolerance region. Further, out of the available 185 (after ….) reanalysis effect sizes (185) 55.64% were inside of this region (Fig. S11). | | With a broader tolerance region (+/- 0.20 Cohen’s d) we found that from the available reanalysis effect sizes r dplyr::filter(ind\_within\_tol\_reg\_20, is\_within\_region == "Outside of tolerance region") |> dplyr::pull(percentage)% (r dplyr::filter(ind\_within\_tol\_reg\_20, is\_within\_region == "Outside of tolerance region") |> dplyr::pull(n)) were outside of the tolerance region. |
| NOT IN MANUSCRIPT | | To investigate the robustness of our tolerance region threshold we calculated the proportion of reanalysis effect size estimates within the tolerance region for a range of thresholds between 0.05 and 0.1. |
| Figure 2c + 2d | | We were interested to see whether these robustness results show a different pattern when inspecting separately by the disciplines of the studies. @fig-effect-region-discipline and @fig-effect-robustness-discipline show that for the major disciplines (>=10 studies). |
| Figure 2e + 2f | | Here (@fig-effect-region-studytype and @fig-effect-robustness-studytype), we were interested to see whether these results show a different pattern when separating them by study type. |
| Figure 2g | | Here (@fig-effect-region-expertise), we were interested to see whether these results show a different pattern when inspecting them along the reported expertise of the co-analysts. |
| Figure 2h | | Here, we were interested to see whether these results show a different pattern when inspecting them along their prior familiarity with the dataset. @fig-effect-region-familiarity shows that for these results. |
| NOT IN MANUSCRIPT | | Here (Table 4), we were interested to see whether these results show a different pattern when inspecting them along their level of confidence with the suitability of the analysis. |
| Figure 2i | | Here, we were interested to see whether these results show a different pattern when considering sample size. The following @fig-samplesize-region shows no remarkable differences between the two categories. |
| NOT IN MANUSCRIPT | | 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 designs. Following the conduct of the present project, Kümpel & Hoffmann (2022) proposed a formal definition of generalized marginal effects (gMEs) which measure is 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 sub-sample of the 100 studies. See @fig-gme for the results of the gME calculation for our sub-sample of studies. |