Title

Investigating the analytical robustness of the social and behavioural sciences

Journal

Nature, 652(8108)

ISSN

0028-0836 1476-4687

Authors

Aczel, Balazs

Szaszi, Barnabas

Clelland, Harry T

et al.

Publication Date

2026-04-01

DOI

10.1038/s41586-025-09844-9

Peer reviewed

Article

Investigating the analytical robustness of the social and behavioural sciences

https://doi.org/10.1038/s41586-025-09844-9
Received: 27 January 2025
Accepted: 31 October 2025
Published online: 1 April 2026

The same dataset can be analysed in different justifiable ways to answer the same research question, potentially challenging the robustness of empirical science[1–3]. In this crowd initiative, we investigated the degree to which research findings in the social and behavioural sciences are contingent on analysts' choices. We examined a stratified random sample of 100 studies published between 2009 and 2018, in hich, for one claim per study, at least five reanalysts independently reanalysed the original data. The statistical appropriateness of the reanalyses was assessed in peer evaluations, and the robustness indicators were inspected along a range of research characteristics and study designs. We found that 34% of the independent reanalyses yielded the same result (within a tolerance region of ±0.05 Cohen's *d*) as the original report; with a four times broader tolerance region, this indicator increased to 57%. Of the reanalyses conducted, 74% reached the same conclusion as the original investigation, 24% yielded no effects or inconclusive results and 2% reported the opposite effect. This exploratory study indicates that the common single-path analyses in social and behavioural research should not be simply assumed to be robust to alternative analyses[4]. Therefore, we recommend the development and use of practices to explore and communicate this neglected source of uncertainty.

Over the past decade, social and behavioural scientists have been striving to enhance the robustness, objectivity, and replicability of their findings through systemic reforms in the conduct and communication of empirical research. Practices such as preregistration[5], registered reports[6], multisite replications[7], analytical reproducibility checks[8,9], and automated result validation techniques[10] have been investigated and recommended to produce robust and replicable findings. An important aspect of robustness has yet to be systematically charted across these sciences: the contingency of the results on researchers' analytical choices.

In a typical research pipeline, the collected empirical data are analysed by a single analyst or team, and the published report presents a conclusion based on one analytical path, occasionally accompanied by a few robustness tests. The peer review process aims to ensure that the analysis approach meets the statistical and field-specific standards. However, this procedure does not systematically ascertain whether justifiable alternative analytical choices could have led to different results. Theories and empirical designs rarely constrain analysts to a single analytical path. Many degrees of freedom exist in how researchers operationalise their variables, process their data, construct their statistical model, select algorithms and software for model estimation, and define their inference criteria; whether they follow frequentist, Bayesian, or likelihoodist analytical approaches; use machine learning or conduct computational modelling to answer the same research question[1,4]. This inherent freedom of the analyst constitutes the so-called *analytical variability* contained within empirical projects, a key component in the robustness of the statistical results. In practical terms, it is the manifested variation among the choices independent scientists consider justified. Fig. 1 lists some sources of analytical variability that can manifest themselves in analysts' statistical results and the conclusions drawn from the results.

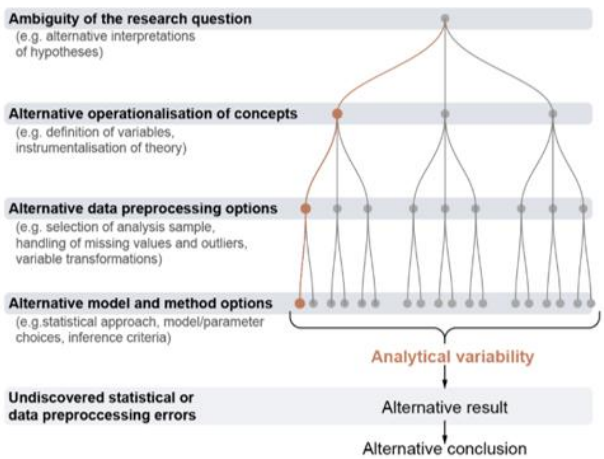

Fig. 1 | Major sources of analytical variability. Analytical variability can arise from the ambiguity of the research question, alternative operationalizations of concepts, variations in data preprocessing options, model and method choices and undiscovered statistical or data processing errors.

One way to explore analytical variability is to employ a *multiverse* methodology[2,11] in which the analyst conducts all combinations of analytic choices they are able to generate across a wide range of reasonable scenarios. Alternatively, in the *multi-analyst* approach, multiple analysts analyse the data following their best judgement. The latter approach requires more organisation, but it takes advantage of alternative expert perspectives without the combinatory expansion of the number of results. A multi-analyst approach also examines naturally occurring variation, empirically answering the counterfactual question of what might have happened if another investigator had considered the same research question using the same data.

Multi-analyst projects[3,12–24] have provided some evidence of the extent to which analysts' individual choices influence the results and conclusions. From economics to neuroscience, these explorations demonstrated that the robustness of empirical findings can be compromised by researcher degrees of freedom[25]. The estimates of previous multi-analyst studies suggest that the variability in effect-size estimates attributable to analytical heterogeneity can exceed the variability one would expect due to sampling error[26].

Do we know how robust published findings are to analytical choices across the social and behavioural sciences? One could argue that multi-analyst projects so far have been purposefully conducted in research areas with little consensus on the best analytical approach or were motivated to demonstrate the potential effect of analytical choices and, therefore, may represent rare cases where alternative analyses produce important differences in results. For example, perhaps the datasets selected afforded greater researcher degrees of freedom than is typical, raising issues of the generalizability of the findings to scientific research more broadly. Differences between academic methodologies and fields also seem plausible - for example, the relatively simple experiments sometimes used in social psychology and behavioural economics may contain fewer analytic decisions than the complex longitudinal observational datasets used in macroeconomics and finance, and thus be more analytically robust in general[22]. To the extent that this is the case, the findings from the existing multi-analyst projects could be biased towards worst-case scenarios, and the traditional analytical practice and review system may not require fundamental adjustments. If, on the other hand, observed results are contingent on the analyst's choices across fields, methodologies, and types of datasets, then the scientific literature could be less robust than is often assumed. If so, the general practices of how we conduct, report, and review empirical analyses should be reformed to address this source of uncertainty.

After conducting 504 re-analyses with the involvement of 457 independent re-analysts on a stratified random sample of 100 social and behavioural studies, we conducted strictly exploratory analyses in order to describe the patterns in the findings. Inspecting the results across different research characteristics and study designs gives rise to a number of hypotheses for future research on how to maximize transparency and address this often-neglected component of scientific uncertainty.

Variability of the results

To explore the robustness of published claims, we selected a key claim from each of our 100 studies, in which the authors provided evidence for a (directional) effect. We presented each empirical claim to at least five analysts along with the original data and asked them to analyse the data to examine the claim, following their best judgement and report only their main result. The analysts were encouraged to analyse those studies where they saw the greatest relevance of their expertise. Therefore, in this study, analytical variability, as a key component of robustness, is defined as the variation among the analytical results when different analysts are provided with the same research questions and the same data.

First, we explored the degree to which the re-analysts produced the same statistics in the re-analysis of each study. We found that in 81% of the studies, the corresponding analysts reported different statistics regarding statistical test families (such as *t*-tests, F-tests, and $\chi 2$ tests) and their values (after rounding them to two decimal places).

A challenge in any multi-analyst project is to find a common metric that allows the results of the different analyses to be compared. A practical solution is to transform the reported point estimates into a standard effect size measure. Although these transformations have limitations and their calculation relies on assumptions that may not hold in all considered analysis settings[25,27–29], for the sake of comparability, we decided to compute Cohen's *d* for each re-analysis, wherever it was feasible. (For an alternative approach, see Additional Results in the Supplementary information.) In our preregistration, we defined that we consider two results qualitatively the same when their effect sizes are within the tolerance region of ±0.05 Cohen's *d*. However, below, we also report analyses with alternative tolerance regions. Our results revealed how far the new estimates were from the original ones (Fig. 2a) and how often the effect sizes of the re-analyses fell within this tolerance region (Fig. 2i).

We found that in 5.26% (5 out of 95) of the studies for which we could obtain the original effect size all re-analysis effect sizes were inside the tolerance region (+/- 0.05 Cohen's *d*) of the result of the original study (Fig. 2a). Out of the 396 available re-analysis effect sizes, 33.84% 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. With a four times broader tolerance region (+/- 0.20

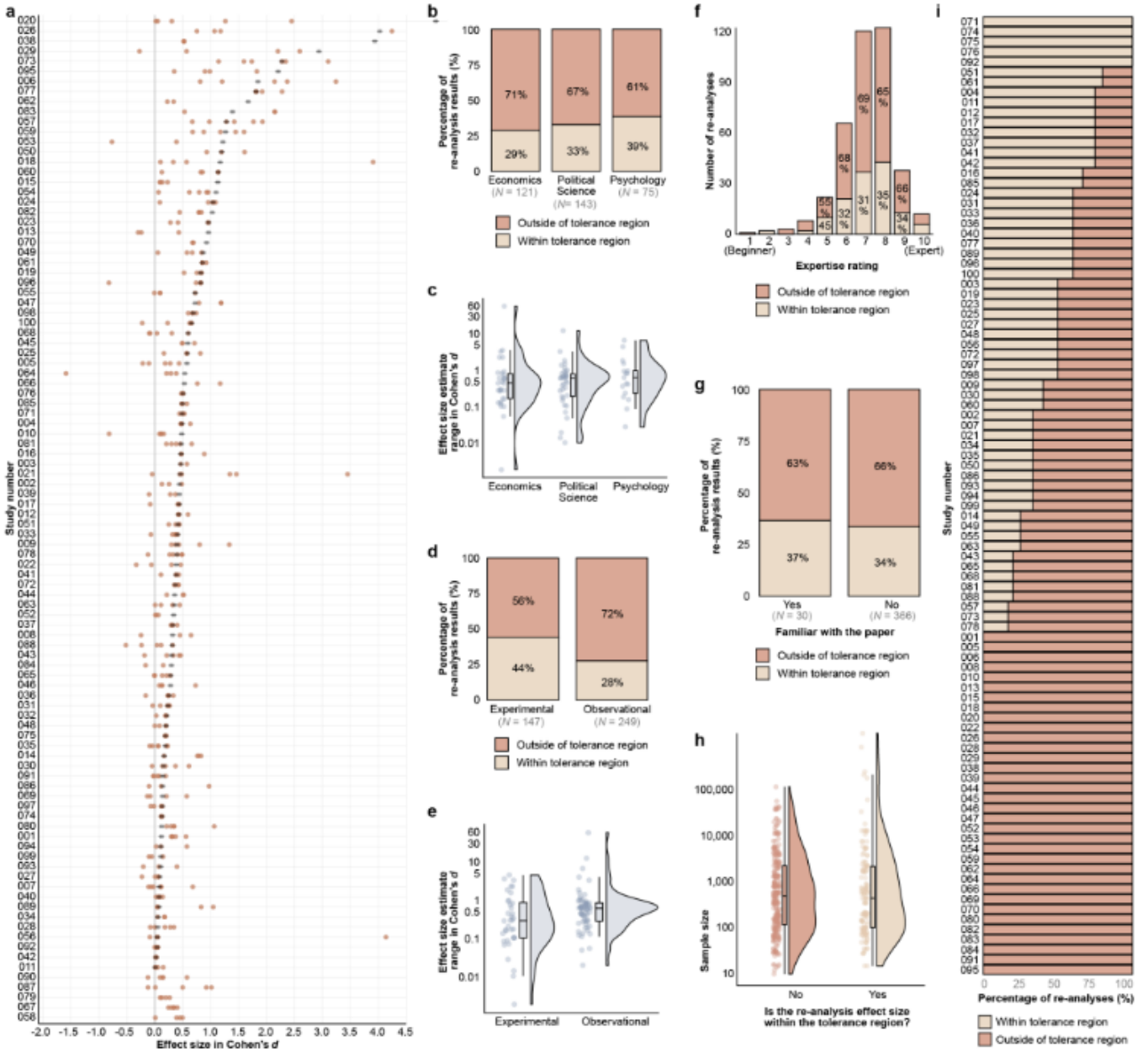

Fig. 2 | Analytical robustness of the statistical results. a, Effect size of the original analysis (grey squares; all represented as positive values) and the effect sizes of the re-analyses (red dots) for each study. The figure displays 415 reanalysis effect-size estimates that were convertible to Cohen's *d* and excludes effect sizes outside the [−2, 4.5] range. For the five studies listed at the bottom of the figure, we could not determine the original effect size because of missing information. Study numbers correspond to the studies listed at https://osf.io/mkwhn. The studies are ordered by the original effect size. **b**, Percentage of reanalysis results falling within or outside 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. **c**, Distributions of effect-size estimate ranges calculated per study for each major discipline. **d**, Proportion of reanalysis results falling within or outside 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. **e**, Distribution of effect-size estimate ranges calculated per study for observational and experimental studies. **f**, Percentage of reanalysis results falling within or outside the tolerance region of the original results of the studies by self-rated expertise (1, beginner; 10, expert). **g**, Percentage of reanalysis results falling within or outside the tolerance region of the original results of the studies by declared familiarity with the study. **h**, Distribution of sample sizes separately for reanalysis effect sizes falling within or outside the tolerance region of the original results. **i**, Proportion of effect sizes falling within the preset tolerance range (+/-}0.05 Cohen's *d*) for each study.

Cohen's *d*), in 23.16% of the studies, all corresponding re-analysis results were inside the tolerance region. Further, out of the available 396 re-analysis effect sizes, 56.57% (224) were inside of this region (Extended Data Fig. 1a).

Alternatively, we could define the tolerance region as the percentage of the given effect size. As an additional robustness test, we varied the tolerance region between +/-5% and +/-20% but it barely made any difference regarding the percentage of robust studies (Extended Data Fig. 1b).

We next considered whether these robustness results vary by the disciplines of the studies, the study designs, the expertise of the analysts, their prior familiarity with the data, and the sample size in the data. Fig. 2b and Fig. 2c show the results for the major disciplines in our sample (>=10 studies). For Fig. 2c, we created an effect-size estimate range for each study as the numerical difference between the highest and lowest estimate of re-analysis effect sizes. In our reading, the listed disciplines do not yield large differences in the robustness of the results. Still, it is reasonable to think that the level of analytical robustness in different disciplines can be influenced by the type of studies that are commonly conducted there. For example, one could conjecture that empirical claims based on observational data show lower robustness of the conclusions since they likely involve more researcher degrees of freedom in terms of viable analysis paths than experimental research settings. Fig. 2d and Fig. 2e explore this question and indicate that the results of studies with observational study designs have lower analytical robustness in our sample, relative to experimental designs (also see Tables S5 and S6).

Considering the analytical variability found in the statistical results of the re-analyses, one immediate concern is that it could be an artefact of a lack of analytical expertise among some re-analysts. Therefore, we explored whether our robustness results exhibit a different pattern when examined in relation to the self-reported statistical expertise of the re-analysts. Visual inspection of Fig. 2f shows no support for this proposition, as a higher level of expertise corresponds with no increase or decrease in the ratio of the reported results being different from the original ones. It is noteworthy, however, that the level of self-perceived expertise was clustered in the higher end of the scale.

Re-analyzing published studies entails a potential risk of bias if the re-analysts' familiarity with a given study influences their choice of analysis. Re-analysts reported that they were familiar with the original study in only 8.13% of cases. Moreover, there was no more than 3% difference in robustness between those who were and those who were not familiar with the original study (Fig. 2g). For both groups, around two-thirds of the estimates fell outside our tolerance region. Finally, we were interested to see whether these robustness results would show a different pattern when considering sample size, as one could assume that studies with larger sample sizes could offer more robust results. Fig. 2h does not support this assumption as the density distributions of the sample sizes for results that are within and outside of the tolerance region are virtually the same. Therefore, studies with large sample sizes are not immune to analytical variability.

We next asked whether the re-analyses show a trend or shift in effect sizes compared to the results of the original studies. If the re-analysis effect sizes randomly vary around the original effect size, we would expect that they are larger or smaller than the original ones with an equal chance. Fig. 3a (re-analysis data trimmed at Cohen's d =< 5) and 3b (Cohen's d =< 1) indicate that re-analysis effect sizes show a tendency to be smaller than the original effect sizes as reflected in their best-fitting (least squares) line. The distribution of original and re-analysis effect sizes also supports this, as the peak of the density distribution of the latter is markedly lower. The mean effect size of the original results is 0.73 (Median = 0.43), whereas for the re-analysis it is 0.49 (Median = 0.35), Cohen's d, computed on ds =< 5. This result is consistent with the possibilities that original authors were biased to report larger effects than re-analysts, that re-analysts were biased to report smaller effects than original analysts, or both.

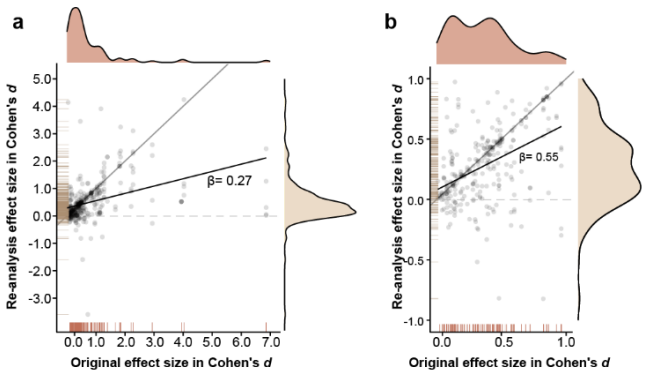

Fig. 3 | Original study effect size versus reanalysis effect size. *The thin diagonal line represents the ideal case in which the reanalysis effect sizes are equal to the original effect size, whereas the thick line shows the best-fitting (least-squares) line of the displayed dots. Density plots of original (n = 95) and reanalysis (n = 504) effect sizes are parallel to their respective axes. β refers to the regression slope.* ***a*** *shows the effect sizes with Cohen's d ≤ 5, whereas* ***b*** *shows the same for effect sizes with Cohen's d ≤ 1.*

Variability of the conclusions

Another focal question of our study was whether the re-analysts reached the same qualitative conclusions as the original study analysts. To answer this question, we asked the re-analysts to implement any statistical re-analysis they deemed most appropriate to test the original claim using the original data, with the goal of arriving at a single conclusion. 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 an effect in the opposite direction as in the original investigation (Fig. 4a).

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. It is important to note, however, that this result is contingent on the level of agreement we use to define analytically robust findings. With a more liberal definition of analytical robustness, this value was 39% when analytical robustness was defined as >80% re-analysis agreement with the original conclusion, and it was 80% when this definition was >50% (the results with alternative levels of agreement are displayed on Fig. 4j).

We examined whether these results show a different pattern when inspecting them along the earlier-mentioned aspects of the analyses. Fig. 4b and Fig. 4c present the proportions of conclusions that were robust in each of the listed disciplines. Just as in the case of analyses of robustness of the statistical results, the listed disciplines do not manifest large differences in robustness of the conclusions, whereas their robustness may be influenced by the study designs most common in a given field or subfield. Fig. 4d supports this notion as it indicates that nearly half of the conclusions from experimental studies remained robust upon independent re-analysis, whereas less than one-third of observational studies yielded robust conclusions. Moreover, Fig. 4e indicates that, although the majority of re-analyses for both study designs reached the same conclusions as the original study, the figure was 13% higher for experimental studies than for observational studies. Just as for the robustness of the results, we can ask whether the deviation from the originally reported claim in terms of conclusions is explained by the re-analysts' lack of analytical expertise. Fig. 4f shows no support for this conjecture when evaluating the pattern of results as a function of self-reported statistical expertise. The same conjecture can be assessed by considering the quality of the submitted statistical analyses that were evaluated by peer evaluators on a subset of the analyses (see Methods). Fig. 4g shows that the proportion of inferentially robust conclusions is numerically larger for analyses that were rated as medium-quality by peer evaluators than for analyses that were rated as high-quality. Whether this pattern was a result of noise or whether more sophisticated analyses are characterized by greater heterogeneity in approaches and results should be the topic of future metascientific projects.

Just as for the analyses of the robustness of the statistical results, we were interested to see whether these results show a different pattern when inspecting them as a function of analysts' prior familiarity with the dataset. Although those familiar with the original study did report the same conclusion in a higher proportion than those who were not familiar, 17.07% of their re-analyses still indicated a conclusion different from the original one (Fig. 4h).

Again, we aimed to explore whether these robustness results would show a different pattern when considering sample size. As presented in Fig. 4i, the density distribution corresponding to the analyses with the different conclusion types shows a comparable spread, suggesting that the conclusions of studies with smaller and larger sample sizes appear to be similarly contingent on analytical choices.

For descriptive information about the re-analysts, peer evaluators, and additional robustness analyses, see Extended Data Figs. 2, 3, 4, and Supplementary Information.

Limitations

This study has a number of limitations. First, our collection of 100 articles represents only a tiny fraction of all the empirical work in the social and behavioural disciplines. Despite our efforts to select a representative sample of published articles across disciplines from the investigated time period, we could not include studies when the underlying data were not obtainable, and we excluded studies when our screening attempt to analytically reproduce the original results following the published procedures failed. We cannot exclude the possibility that these prerequisites, in addition to the self-selection of the analysts, led to sampling bias.

Although we conducted more than 500 analyses, our project included only five independent analyses for most datasets, therefore, we do not know to what degree these analyses capture the full variability of analyses and results for the given research question and dataset. Also, since we re-analysed already published studies and the re-analysts were provided with these studies, the original analysis pipeline could have anchored some of the choices of the re-analysts. On the other hand, some analysts could have been motivated to produce alternative results, as it is a basic incentive of scientists to say something new.

While Cohen's *d* has the advantage of being easy to compute and comparable across different analyses, Kümpel and Hoffmann[39] recently proposed the concept of generalised marginal effects (gMEs), an effect size metric that is both formally applicable and comparable across different statistical models. We had not originally planned to calculate standardised gMEs, and, accordingly, did not collect all required analysis outputs to compute them across the board. Still, we calculated gMEs for a sample of our studies to showcase their potential for future multi-analyst studies (Supplementary Fig. 1).

We presented some exploratory analyses, but there are many other factors to explore that could contribute to analytical variability (e.g., topical expertise). Finally, despite our best efforts to conduct quality checks on the re-analyses to ensure the soundness of the analytic strategies[16], it is possible that some of the discrepancies between the original and the new

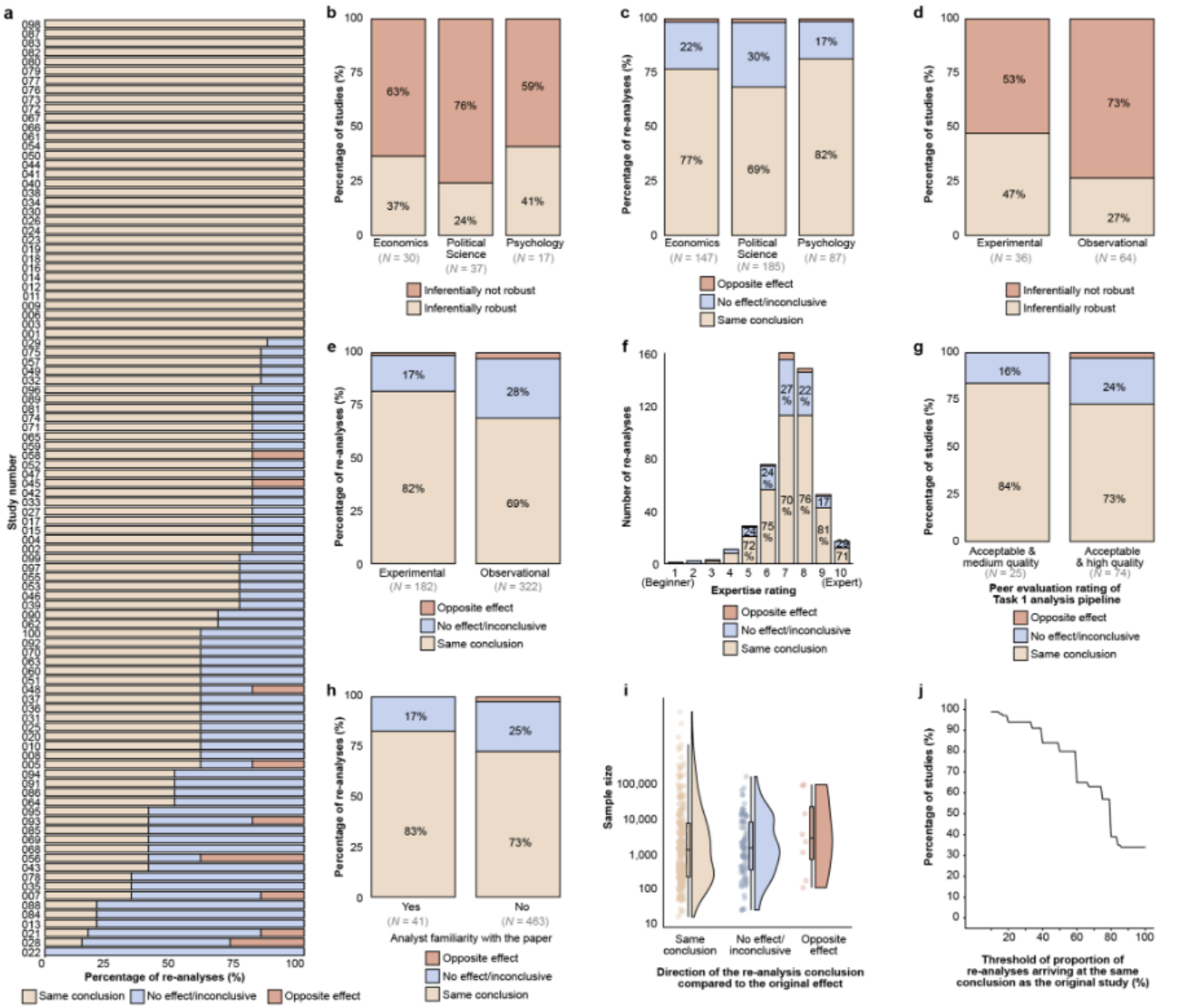

Fig. 4 | Analytical robustness of the conclusions.

a, Proportion of same conclusion, no effect/inconclusive results, and opposite direction conclusions for each study. Study numbers correspond to studies listed in https://osf.io/mkwhn. **b,** 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). **c,** 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. **d,** 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). **f,** 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)). **g,** 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. **h,** Proportion of same conclusion, no effect/inconclusive, and opposite effect of the re-analyses by declared familiarity with the study. **i,** 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. **j,** Percentage of studies with robust conclusions above different levels of re-analysis consensus. Re-analysis consensus refers to the agreement among the conclusions drawn by the original study and the independent re-analyses.

results are due to weaknesses in the re-analysts' approach rather than equally justifiable alternative analysis decisions. It is likewise possible that there are weaknesses in the original analysts' approaches. It is unknown whether the quality control processes for the re-analysts resulted in better, worse, or similar overall quality of analysis decisions as compared with the quality control processes for original analysts' decisions. The declared statistical expertise of the re-analysts makes us believe that the observed heterogeneity in analyses and outcomes is a good representation of variation in informed analysis decision-making in social-behavioural research.

Discussion

Are published results in the social and behavioural sciences robust to independent re-analyses? The present exploration shows considerable variability due to researcher degrees of freedom in statistical choices. 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. Notably, the new conclusions converged with the original ones in 74% of the individual re-analyses. Our descriptive results suggest a number of hypotheses concerning the circumstances in which we could expect greater analytical variability.

Why can there be multiple answers?

Faced with the variability in the analysts' effect-size estimates and conclusions, one intuitive hypothesis is that the variation must be due to researcher characteristics, such as statistical or field-specific knowledge. Previous multi-analyst studies found little to no effect of researcher-specific characteristics, such as experience in the field or statistical expertise[16,19,28]. Instead, they suggest that analytic results are dependent on the particular choices that the analysts make among similarly acceptable data processing and analysis choices[28]. For example, when 46 independent analyst teams analysed the same speech dataset to answer the same research question, the authors concluded that "depending on the choice of how the speech signal is operationalised, researchers might find evidence for or against a theoretically relevant prediction" (p. 21)[28].

In line with previous findings, our results showed no strikingly different patterns across self-reported statistical expertise and experience in a matching field (see Supplementary Information). At the same time, the few analysts who reported that they were familiar with the original article produced alternative results and conclusions at a comparable rate. More importantly, our peer evaluation process did not indicate that the analytical variability of the re-analyses was due to inadequate statistical practices. These results are in line with Menkveld et al.[22] in which the quality assessment of the proposed analysis pipelines did not statistically explain the results.

Another line of thought would suggest that the lack of robustness in the original published results reflects some conceptual ambiguity in the theories or methodology[30]. Research hypotheses are often short verbal expressions that do not force the specifications of the analyses. The underspecification of claims[31] could represent a major source of ambiguity in analytic decisions. We could not test the role of hypothesis ambiguity in a controlled manner, but it is a plausible contributor considering that social science theories often make general claims across many variables, creating theory-laden choice points regarding how constructs are operationalised, and how hypotheses are tested[32].

Regarding methods, we explored our results by separating them by experimental and observational study designs, and observed that the proportions of results and conclusions that were analytically robust were 15-20% higher for the experimental studies. The estimated range of effect sizes was also apparently wider for observational studies compared to experimental ones. This exploratory finding motivates the hypothesis that the increased control over data collection circumstances and the reduced number of variables in experimental versus observational research translate to more limited analytic flexibility. Notably, however, there was still substantial statistical variability among findings from experimental studies.

Why do these findings matter?

Where multiple acceptable analytical paths exist, researchers can use this freedom opportunistically[33,34] and bias the results towards desired findings ("myside bias"[35]). The much-discussed credibility challenges in the social and behavioural sciences stem partly from the suspicion that the prevailing incentive systems for publication encourage researchers to report and interpret empirical data to serve non-epistemic goals such as storytelling[36]. Reform initiatives, such as the preregistration of research and analysis plans, aim to decrease researcher degrees of freedom to tweak the analytic method or the research question to the observed data. Would results in these fields become markedly more credible if every study was preregistered? Since preregistration is a protection against overfitting, we hypothesize that it would reduce or eliminate the observed finding that original analyses showed stronger evidence for positive results than re-analyses. However, we also hypothesize that preregistration would have little impact on the observed heterogeneity across alternative analysis strategies since registering and following a single analytic path constrains the analysts only from choosing opportunistically from the alternative analytical paths. Still, it does not confer any unique

statistical or epistemic status to the pre-selected analytic path[27]. Unexplored but alternative justifiable analyses applied to the same data could still lead to very different results. The present exploration is clear about the presence of this variability in approaches, results, and inferences in the social and behavioural sciences. Without exploring this variability, authors cannot guarantee consumers of their research that the reported conclusions hold a privileged status over alternative conclusions.

What can we do?

The outcomes of this project suggest that the empirical answers to research questions in the social and behavioural sciences depend on the analytic paths taken to pursue them. Therefore, we advocate for the broader adoption of approaches that explore, recognise, and address the uncertainty created by analytical variability.

Two main types of solutions are (1) multi-analyst studies, such as our own, where multiple investigators independently follow their own approach, and (2) the multiverse[2,11,37] approach, where one investigator or team performs numerous analyses across the set of reasonable pipelines. Conducting exploratory studies to identify analytical uncertainties and holding out samples are further advisable practices to tackle analytical variability.

Project leaders aiming to conduct multi-analyst studies can consult various tutorial papers and guidelines. Aczel et al.[38] provide an expert consensus guideline on the entire life-cycle of multi-analyst projects from recruiting suitable analysts, through conducting the project to the reporting of the outcomes. Kümpel & Hoffmann[39] offer a framework for synthesising objective outcome metrics. The Subjective Evidence Evaluation Survey[40] is a tool for systematically exploring and quantifying subjective measures of evidence in multi-analyst studies, allowing analysis teams to subjectively reflect on various aspects of evidence, such as coherence, robustness, and relevance, as well as the quality of the research design and data. Multiverse analysis is also useful, especially when the dataset cannot be shared with other research groups due to confidentiality reasons or when there are insufficient human resources to recruit several independent analysts. Several guideline papers help researchers conduct and interpret such analyses[2,37,41–43].

Recently, many scholars have called for a stronger focus on replication in science[44]. Similar to preregistration, however, replications are unlikely to help address the robustness of results to multiple analysis strategies as they intentionally repeat the same (or at least a very similar) analysis path. In this sense, replications can help detect bodies of work in which authors may have leveraged their researcher degrees of freedom to generate results that are in line with their own or the journal's expectations. All other things being equal, a severely *p*-hacked literature should contain fewer replicable findings. And yet, replicability does not eliminate analytical variability itself. Nevertheless, having multiple studies creates an opportunity to observe if analytical variability is, itself, replicable. For example, imagine that Study A provides evidence for a claim with Analysis 1 but not with Analysis 2. If several replications also find evidence for the claim with Analysis 1 but not with Analysis 2, then the analytic choices are directly implicated in how evidence for the phenomenon is observed. However, if it is random across replications whether Analysis 1 or Analysis 2 provides evidence for the claim, then the implications of the analytic variability are very different. The combination of replications and robustness investigations will facilitate the advancement of stronger theoretical underpinnings of the topics of study, and could reduce analytical variability in the long run by creating a more direct mapping between theory and measurement[30,45,11].

All in all, we argue that the scholarly communication system could foster more engagement with systematic and transparent robustness testing. As a starting point, the research data shared openly alongside codebooks and analysis scripts is a prerequisite for any assessment of analytical robustness. Research findings of particular scientific or societal importance could be accompanied by robustness reports[46] that summarise the results of alternative theory-motivated analytic choices by independent analysts. This publication format already provides a platform for analysts to scrutinise the fragility of the findings before they have a major impact on scholarship and policy.

What did we learn about the robustness?

Our results support the view that the results in social and behavioural science studies are contingent on the analyst's choices, and if an analyst reports a single result from a single analytical path, they have not exhausted the possible answers that the dataset can provide. This finding aligns with the conclusions drawn by Wagenmakers, Sarafoglou, and Aczel[4], that the belief that "for any dataset, there exists a single, uniquely appropriate analysis procedure" and "multiple plausible analyses would reliably yield similar conclusions" (p. 424) are no more than statistical myths. Without multi-analyst and multiverse approaches, the fragility of empirical findings remains.

Nonetheless, we emphasise that an optimistic or pessimistic interpretation is a matter of perspective and greatly depends on what evidential support we expect from a given study. Therefore, whether a result is satisfyingly robust will always depend on our epistemic needs and the precision we expect from our results. We caution against using blanket rules in aggregating or interpreting results across different analytical approaches within the same investigation.

Objectivity is a fundamental ideal of science, implying that claims about the world should not be contingent on the predispositions of the claimant. What our results reveal is not that we must distrust or reject the results of the past, including the studies we analysed. Instead, they suggest that we should adopt greater caution about the evidence that single analytical paths can offer to support social and behavioural science claims. We believe that the limitations of "single-shot" analyses cut across numerous scientific disciplines. Methodological innovations, such as multi-lab collaborations, multi-analyst approaches, or multiverse methods, could increase the robustness of the social and behavioural sciences, and perhaps more broadly, in other empirical fields.

Methods

All methods and procedures in this study were vetted by a panel of experts with prior experience in multi-analyst studies or who are specialists in the relevant methodology (see Supplementary Information).

Preregistration

The methods, materials, analysis plan, peer evaluation, and data management strategy of the project were preregistered on the OSF. Deviations from the registered plan are reported and explained in the "Deviations from preregistration" supplementary document.

Ethical considerations

The datasets resulting from this project were not considered human subject research and are covered under an umbrella ethics protocol that was managed by the Center for Open Science (COS) (BRANY SBER IRB protocol #21-056-749), with concurrence from the United States Naval Information Warfare Center Pacific, HRPO. The institutional ethics board of the Faculty of Education and Psychology at Eötvös Loránd University, Budapest, Hungary, determined that the re-analysts are not considered research participants and that the project raises no ethical concerns.

Materials

Selection of studies

The selection of studies was completed in two stages. In the first stage, the SCORE team created an initial study and claim collection. From this collection, we selected our sample using additional criteria.

In the SCORE project, a stratified random sample of 600 articles was identified from a larger pool of randomly stratified ~30,000 articles from 62 journals, published between 2009 and 2018. The journals covered the main branches of social and behavioural sciences (criminology, economics, education, health-related, marketing/organisational behaviour, management, political science, psychology, public administration, sociology). To obtain the original studies, the following steps were taken: First, the paper was reviewed. If data and/or code were available, they were downloaded and saved into a project on OSF. If data and/or code were not available, the SCORE team attempted to contact the corresponding author to request that they share the data and code used for the original publication. Studies were excluded from the sample if they did not contain at least one inferential test using non-simulated, human data, where human data are defined at any level of human organisation (e.g., the individual person, family, political entity, firm, economic unit). The majority of the studies were tested for analytic *reproducibility* using the original specification, which is to be distinguished from *robustness* to alternative specifications. 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). If the analytic reproduction failed, the paper was removed from the pool. Therefore, the present project focused solely on robustness to alternative specifications and did not conduct direct reproducibility checks using the original specification, as these had already been carried out by SCORE. Further details of the SCORE methodology (list of journals, selection process, etc.) are available in the original report[45].

In the present work, a further requirement of the selected studies was to contain a single inferential statistical test result that corresponded to the claim with our instructions. Thus, we ensured that given the claim and the instructions, no other statistical result could correspond to the claim in the original article. If all potential claims from the study were too ambiguous and, therefore, could not be linked with a single inferential test statistic with the specification instructions, the study was excluded from our sample. The above-described study selection process was continued until we reached our target number of 100 studies, corresponding claims, and datasets.

The selected studies and all available corresponding data and materials were made available to the re-analysts so that they could fully understand the selected claim and approach. There are trade-offs for how much information to give to the re-analysts to conduct re-analyses. Complete blinding of the original analysis strategy would ensure an entirely independent decision-making process about how to analyse the data.

However, in much scientific writing, there is insufficient clarity in the description of the theoretical background, rationale, and specification of the conceptual model to be tested. In some papers, there is a clean break between these and clear hypotheses to test. In other papers, the narrative intermixes theoretical statements and analysis decisions and may not clearly state hypotheses or how they correspond with observed results. As a consequence, attempts to blind papers inevitably lead to variation in what is blinded across papers and many subjective decisions about what should be blinded (because it provides information about analysis strategy) and what can remain unblinded (because it provides information about theory and rationale). A major risk of those blinding decisions is that important information could be removed, which would weaken the re-analysts' ability to conduct a fair re-analysis of the original claim. As such, we opted for complete transparency of the original article so that no potentially important information was missing for the re-analysts, and we instructed re-analysts that they should create an analysis plan based on their own decisions for how best to assess the study's claim. On balance, this increases the risk of dependent decision-making but reduces the risk of misspecification of the hypothesis and rationale of the original research. In this context, we judged the latter to be a more important precondition for conducting an informative study.

Claim selection

Claim selection was built on Phase 1 of the SCORE project effort. The claims identified for Phase 1 of SCORE were executed according to a "single trace" approach, where only a single claim trace was extracted from the article, which corresponded to one statistically significant inferential test result. Within the current project, first, the lead team ensured that the extractions (i) are understandable, (ii) contain only one claim, (iii) indicate the direction of the effect, (iv) there is a statistical hypothesis test-based result provided in the article that corresponds to the claim; and (v) the claim was phrased on a conceptual and not statistical level. If not, then they extracted the part of the claim that is relevant, or if this could not be achieved, they selected another more suitable sentence from the abstract, or if this could not be achieved, they searched for another suitable sentence from other parts of the article that could satisfy all of our criteria. When none of these steps presented a claim that satisfied the expectations, then the given article was not used in our study (for their list and explanation of dropping, see our data table). Where an expression of a claim has been judged by the lead team as ambiguous or rhetorical, they substituted the expression with an ellipsis mark (e.g., "dramatically increased" to " ... increased") while preserving the original wording and the meaning of the claim. Only in cases where, due to the selection, the wording of the claim became complicated, ungrammatical, or contained an ambiguous definition or an unexplained abbreviation, did the core team make necessary (and marked) adjustments in the grammar or wording of the claim while preserving the original meaning of the extraction. For example, the following selection "Three factors increase the salience of the proliferation threat: (1) prior violent militarized conflict..." was changed to "[prior violent militarized conflict] increase[s] the salience of the proliferation threat ...". The list of claims can be found at https://osf.io/mkwhn.

Analysis instructions

For the re-analysts' second task, instructions were needed in cases where the original paper contained more than one statistical analysis corresponding to the high-level claim, in order to be able to compare the new result to the one in the original paper. For this, the lead team prepared certain instructions (e.g., data selection, exclusions) that single out only one statistical result in the original paper. The instructions always remained circumstantial (e.g., data selection, exclusions, choice of measurement) and never gave direct instructions to the choice of statistical approach or full specification of the model.

Procedures

Re-analyst recruitment

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

Participation in the project was advertised on social media, at conferences, in mailing lists (e.g., SCORE collaborator list), via personal networks, and in research newsletters. As a response to our recruitment call, 1141 researchers signed up to participate in our study. Out of these volunteers, 459 signed up to analyse at least one dataset and submitted their work by the deadline or an extended deadline. From all the eligible volunteers, we selected re-analysts and peer evaluators on a first-come, first-served basis. The expectation of participation in the study was experience with conducting statistical analyses, and this was communicated to the volunteers from the start of the recruitment. Re-analysts were informed that they would qualify as authors on the publication of this study if (1) they completed their analyses and submitted all required materials and the post-analysis survey on time; (2) their analyses passed the peer evaluation, and (3) they reviewed and approved the manuscript in time.

Re-analysts received a flat fee of 100 USD for each of their completed re-analyses (including both Task 1 and Task 2) if they submitted their work before March 2023, the deadline of the

grant budget, unless they were from an embargoed country, in which case we were unable to transfer any payment. Peer evaluators received a flat fee of 10 USD per peer evaluation. Any further volunteers were informed that this payment did not apply to them.

Upon joining the project, the volunteers for re-analysis were required to accept the project requirements. They were informed about (a) their tasks and responsibilities; (b) the project confidentiality agreements; (c) the plans for publishing the research report and presenting the data, analyses, and conclusion; (d) the conditions for an analysis to be included or excluded from the study; (e) that their names will be publicly linked to the analyses; (f) the re-analysts' rights to update or revise their analyses; (g) the project time schedule; and (h) the nature and criteria of compensation. Re-analysts were informed that, whereas they could consult other researchers during their analyses, they could not work in teams in this project. Before discussing the details of the analyses with others, the re-analysts were asked to ascertain that the person was not another re-analyst on that dataset. All communication materials of this study are openly available on the public repository of the project at https://osf.io/nvy8a.

Assignment of analyses and tasks

The following procedure was first piloted with two analysts to learn about the practical challenges and time demands of the following tasks. As the results of those analyses were not of central interest, we kept no records of them.

First, each re-analyst was asked to assign themselves to one study, but at later rounds of recruitment, we allowed re-analysts to complete analysis on another paper, other than the one they completed earlier. They were asked to choose those studies where they saw the greatest relevance of their expertise. The authors of the original study could not be the re-analysts of that study.

For several practical reasons, the re-analyses were not started at the same time for each study and each analyst. Firstly, it took us several rounds of recruitment to gather the target number of analyses for each study, mainly due to dropouts, delays, unplanned personal difficulties, and a shortage of staff. Secondly, our analysts found it difficult to retrieve, open, or interpret some of the datasets. In some cases, we had to reach out to the original authors, causing further delays in the project. The task of the re-analysts was to reflect on the corresponding claim (see claim selection) by re-analyzing the corresponding data. The re-analysts were provided with access to the datasets, extracted claims, the original articles, and all the corresponding materials. They were informed that their analyses should be conducted preferably with scripts that could reproduce all their results (including data preprocessing, extraction of test statistics and p-values/Bayes Factors, computing effect-size measures, etc.), but they could use the statistical software of their choice to produce an analysis script. Re-analysts were asked to write and structure their code such that others could understand their analysis scripts (e.g., by annotating the different analysis steps), and they were also informed that the analysis scripts from all analysts would be made publicly available with their names linked to the analyses.

Re-analysts received two main tasks for each study, where Task 2 was given after the completion of Task 1. Once Task 1 was submitted, the analysts could not change the submission of Task 1 unless they were asked by the lead team to provide some missing information from their analysis.

Task 1 The re-analysts were asked to reflect on the selected claim by re-analyzing the corresponding data. They could conduct and report as many analyses as they wished, but they had to draw a single conclusion from their analysis. They were asked to report their analyses and indicate whether their results provided evidence for the relationship/effect as claimed by the original study.

Task 2 For this task, the re-analysts had to produce only one statistical result corresponding to the claim they studied in Task 1, which would be compared to a statistical result in the original paper. The lead team provided certain instructions (e.g., data selection, exclusions) for this analysis to be able to compare the new result to one result in the original paper (see Analysis instructions section). Re-analysts were asked to report their results in terms of statistical families of r, z-, t-, F-, or χ^2 tests (or their non-parametric versions). In addition, they were asked to report sample sizes (e.g., per group) and the corresponding degrees of freedom. By this means, most results could be translated into standardized coefficients by the coordinators.

The reason for requiring two analyses from the re-analysts was that they served two different aims. The results of Task 1 aimed to answer our first preregistered project question: "Do different analysts arrive at the same conclusions as the analyst of the original study?", whereas the results of Task 2 aimed to answer our second preregistered project question: "Do different analysts arrive at the same effect estimates as the analyst of the original study?" We found that asking only one of the tasks would not have been sufficient to fully address both questions. In Task 1, researchers were not constrained to one analysis, so they could have produced more than one statistical result in order to draw a conclusion from the dataset. Therefore, in Task 1, it was not guaranteed that we would be able to select a single effect size from each analyst in order to answer our second project question. Another challenge to finding an answer to our second question was that in some of the original articles, one claim could have had more than one corresponding statistical result listed. In these cases, we prepared instructions for Task

2 in order to single out only one statistical result in the original paper. For example, if the original study contained two corresponding regression models, one with some exclusions and one with no exclusions, then we chose one of them (e.g., the latter), and instructed the re-analysts not to apply any exclusions to the analysed data. In all other regards, re-analysts were free to conduct their calculations according to their best judgment.

After completing the analysis and writing up the methods, results, and conclusion, re-analysts were expected to upload their analysis code (if available) to the corresponding OSF folder. Their reported methods, results, and conclusions were collected via an online form (see https://osf.io/fjnhz/). When uploading the materials, they were also asked to fill out a post-analysis survey. All major communications between the core project team and re-analysts from the study are openly available on the public repository of the project.

Peer evaluations

The goal of peer evaluation in this project was to assess whether the applied analytical choices are acceptable and whether the reported conclusion follows from the statistical results. By acceptable, we mean that peer evaluators agree that the analysis pipeline is within the variations that could be considered appropriate by the scientific community in addressing the given analytical task.

The peer evaluation phase did not address potential errors in translating the description of the analytic methodology into analysis scripts. To mitigate potential gross errors in the analysis, peer evaluators were provided with a thorough and standardised description of the results and conclusions obtained using the described analysis, including sample sizes, the effect size, the test statistic, and degrees of freedom. From the description of the dataset, the description of the analysis, and the reported results and conclusions, peer evaluators were able to identify potential flaws in the implementation of the analysis that could stem from errors and/or mismatches.

Assignment of the analyses

When assigning the volunteer peer evaluators to analyses, the initial rule was that they should not evaluate any re-analyses conducted on datasets they had re-analysed as a re-analyst. In practice, for logistical reasons, this rule was applied in all but six cases (i.e., 98.69% of peer evaluations were carried out on a dataset that was different from the dataset they analysed themselves). They were asked to choose to evaluate those analyses where they see the greatest relevance of their expertise. If, after choosing a study to evaluate, a peer evaluator did not feel sufficiently skilled/experienced to judge whether the proposed analysis was acceptable, he/she was told not to fill out our template and should return the re-analysis to the pool and choose a new one.

Peer Evaluation Procedure

For details, see the corresponding section in the Supplementary Information.

Analysis methods

This exploratory study contains no inferential statistics. Besides the frequency- and proportion-based summary statistics, we calculated only the effect sizes of the results from the original articles and the re-analyses.

Cohen's d effect sizes

Following our preregistration, we converted all results into Cohen's *d*s wherever possible. For a number of cases, we could not achieve this due to missing information in the original studies or reported statistics that cannot be converted into Cohen's *d* (e.g., logistic regression). All the conversions are listed in the R scripts and the data documentation. All the original effect sizes are listed as positive values, and the re-analysis effect sizes are negative only when they showed an opposite effect compared to the original study.

For further information on methods, see Supplementary Information.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Study data and materials are available on the project OSF (https://osf.io/q5h2c/) and GitHub repositories (https://github.com/marton-balazs-kovacs/multi100/). Archived data include the original datasets or a description of how to gain access to them. Our shared materials include all the survey questions and the general communication texts and instructions that we sent to the re-analysts and peer-evaluators. We excluded from our data files the email addresses of the re-analysts, as well as the records of those analysts who did not comply with the instructions and did not submit all the required analyses by the deadline. For further details about our exclusion criteria and procedure, see our Supplementary Information document.

Code availability

All analysis codes for this project are available at https://github.com/marton-balazs-kovacs/multi100.

References

1. Silberzahn, R. & Uhlmann, E. L. Crowdsourced research: many hands make tight work. *Nature* **526**, 189–191 (2015).
2. Patel, C. J., Burford, B. & Ioannidis, J. P. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J. Clin. Epidemiol.* **68**, 1046–1058 (2015).
3. Botvinik-Nezer, R. et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020).
4. Wagenmakers, E.-J., Sarafoglou, A. & Aczel, B. One statistical analysis must not rule them all. *Nature* **605**, 423–425 (2022).
5. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl Acad. Sci. USA* **115**, 2600–2606 (2018).
6. Chambers, C. What's next for registered reports? *Nature* **573**, 187–189 (2019).
7. Coles, N. A., Hamlin, J. K., Sullivan, L. L., Parker, T. H. & Altschul, D. Build up big-team science. *Nature* **601**, 505–507 (2022).
8. Brodeur, A., Dreber, A., Hoces de la Guardia, F. & Miguel, E. Replication games: how to make reproducibility research more systematic. *Nature* **621**, 684–686 (2023).
9. Brodeur, A., Mikola, D. & Cook, N. Mass reproducibility and replicability: a new hope. *SSRN Electron. J.* https://doi.org/10.2139/ssrn.4790780 (2024).
10. Nuijten, M. B. & Polanin, J. R. "statcheck": automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Res. Synth. Methods* **11**, 574–579 (2020).
11. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.* **11**, 702–712 (2016).
12. Bastiaansen, J. A. et al. Time to get personal? The impact of researchers' choices on the selection of treatment targets using the experience sampling methodology. *J. Psychosom. Res.* **137**, 110211 (2020).
13. Hoogeveen, S. et al. A many-analysts approach to the relation between religiosity and wellbeing. *Religion Brain Behav.* **13**, 237–283 (2022).
14. Salganik, M. J. et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc. Natl Acad. Sci. USA* **117**, 8398–8403 (2020).
15. Schweinsberg, M. et al. Same data, different conclusions: radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organ. Behav. Hum. Decis. Process.* **165**, 228–249 (2021).
16. Silberzahn, R. et al. Many analysts, one data set: making transparent how variations in analytic choices affect results. *Adv. Methods Pract. Psychol. Sci.* **1**, 337–356 (2018).
17. Starns, J. J. et al. Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. *Adv. Methods Pract. Psychol. Sci.* **2**, 335–349 (2019).
18. Veronese, M. et al. Reproducibility of findings in modern PET neuroimaging: insight from the NRM2018 grand challenge. *J. Cereb. Blood Flow Metab.* **41**, 2778–2796 (2021).
19. Breznau, N. et al. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proc. Natl Acad. Sci. USA* **119**, e2203150119 (2022).
20. Huntington-Klein, N. et al. The influence of hidden researcher decisions in applied microeconomics. *Econ. Inq.* **59**, 944–960 (2021).
21. Trübutschek, D. et al. EEGManyPipelines: a large-scale, grassroots multi-analyst study of electroencephalography analysis practices in the wild. *J. Cogn. Neurosci.* **36**, 217–224 (2024).
22. Menkveld, A. J. et al. Nonstandard errors. *J. Finance* **79**, 2339–2390 (2024).
23. Sarstedt, M. et al. Same model, same data, but different outcomes: evaluating the impact of method choices in structural equation modeling. *J. Prod. Innovation Manage.* **41**, 1100–1117 (2024).
24. Schilling, K. G. et al. Tractography dissection variability: what happens when 42 groups dissect 14 white matter bundles on the same dataset? *Neuroimage* **243**, 118502 (2021).
25. Gelman, A. & Loken, E. *The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No "Fishing Expedition" or "p-Hacking" and the Research Hypothesis Was Posited Ahead of Tim* https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf (Columbia Univ., 2013).
26. Holzmeister, F. et al. Heterogeneity in effect size estimates. *Proc. Natl Acad. Sci. USA* **121**, e2403490121 (2024).
27. Coretta, S. et al. Multidimensional signals and analytic flexibility: estimating degrees of freedom in human-speech analyses. *Adv. Methods Pract. Psychol. Sci.* **6**, 25152459231162567 (2023).
28. van Assen, M. A., Stoevenbelt, A. H. & van Aert, R. C. The end justifies all means: questionable conversion of different effect sizes to a common effect size measure. *Religion Brain Behav.* **13**, 345–347 (2023).
29. Kümpel, H. & Hoffmann, S. A formal framework for generalized reporting methods in parametric settings. Preprint at https://arxiv.org/abs/2211.02621 (2022).
30. Auspurg, K. & Brüderl, J. Has the credibility of the social sciences been credibly destroyed? Reanalyzing the "Many Analysts, One Data Set" project. *Socius* **7**, 23780231211024421 (2021).
31. Scheel, A. M. Why most psychological research findings are not even wrong. *Infant Child Dev.* **31**, e2295 (2022).
32. Oberauer, K. & Lewandowsky, S. Addressing the theory crisis in psychology. *Psychon. Bull. Rev.* **26**, 1596–1618 (2019).
33. Brodeur, A., Cook, N. & Heyes, A. Methods matter: p-hacking and publication bias in causal analysis in economics. *Am. Econ. Rev.* **110**, 3634–3660 (2020).
34. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
35. Stanovich, K. E. *The Bias That Divides Us: The Science and Politics of Myside Thinking* (MIT, 2021).
36. Gernsbacher, M. A. Rewarding research transparency. *Trends Cogn. Sci.* **22**, 953–956 (2018).
37. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve analysis. *Nat. Hum. Behav.* **4**, 1208–1214 (2020).
38. Aczel, B. et al. Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife* **10**, e72185 (2021).
39. Sarafoglou, A. et al. Subjective evidence evaluation survey for multi-analysts studies. *R. Soc. Open Sci.* **11**, 240125 (2024).

40. Del Giudice, M. & Gangestad, S. W. A traveler's guide to the multiverse: promises, pitfalls, and a framework for the evaluation of analytic decisions. *Adv. Methods Pract. Psychol. Sci.* **4**, 2515245920954925 (2021).
41. Liu, Y., Kale, A., Althoff, T. & Heer, J. Boba: authoring and visualizing multiverse analyses. IEEE Trans. Visual Comput. Graphics 27, 1753–1763 (2020).
42. Olsson-Collentine, A., van Aert, R. C. M., Bakker, M. & Wicherts, J. Meta-analyzing the multiverse: a peek under the hood of selective reporting. Psychol. Methods 30, 441–461 (2025).
43. Zwaan, R. A., Etz, A., Lucas, R. E. & Donnellan, M. B. Making replication mainstream. Behav. Brain Sci. 41, e120 (2017).
44. Bartoš, F. et al. Introducing synchronous robustness reports. Nat. Hum. Behav. 9, 635–637 (2025).