Title: Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty

Authors: Nate Breznau[1]*†§, Eike Mark Rinke[2]†, Alexander Wuttke[3]†, Hung H.V. Nguyen[1,4], Muna Adem[21], Jule Adriaans[10], Amalia Alvarez-Benjumea[35], Henrik K. Andersen[6], Daniel Auer[37], Flavio Azevedo[62], Oke Bahnsen[37], Dave Balzer[24], Gerrit Bauer[32], Paul C. Bauer[37], Markus Baumann[17], Sharon Baute[56], Verena Benoit[57,32], Julian Bernauer[37], Carl Berning[24], Anna Berthold[57], Felix S. Bethke[42], Thomas Biegert[33], Katharina Blinzler[11], Johannes N. Blumenberg[11], Licia Bobzien[18], Andrea Bohman[52], Thijs Bol[56], Amie Bostic[50], Zuzanna Brzozowska[88,85], Katharina Burgdorf[37], Kaspar Burger[92,55], Kathrin Busch[20], Juan Carlos-Castillo[61], Nathan Chan[49], Pablo Christmann[11], Roxanne Connelly[91], Christian S. Czymara[13], Elena Damian[27], Alejandro Ecker[37], Achim Edelmann[47], Maureen A. Eger[52], Simon Ellerbrock[37], Anna Forke[48], Andrea Forster[56], Chris Gaasendam[27], Konstantin Gavras[37], Vernon Gayle[91], Theresa Gessler[92], Timo Gnambs[29], Amélie Godefroidt[40], Max Grömping[14], Martin Groß[82], Stefan Gruber[59], Tobias Gummer[11], Andreas Hadjar[74], Jan Paul Heisig[73], Sebastian Hellmeier[69], Stefanie Heyne[37], Magdalena Hirsch[73], Mikael Hjerm[52], Oshrat Hochman[11], Andreas Hövermann[15], Sophia Hunger[73], Christian Hunkler[19], Nora Huth[84], Zsófia S. Ignácz[13], Laura Jacobs[53], Jannes Jacobsen[9,5,10], Bastian Jaeger[51], Sebastian Jungkunz[64,57,65], Nils Jungmann[11], Mathias Kauff[36], Manuel Kleinert[25], Julia Klinger[62], Jan-Philipp Kolb[7], Marta Kołczyńska[43], John Kuk[79], Katharina Kunißen[75], Dafina Kurti Sinatra[20], Alexander Langenkamp[13], Philipp M. Lersch[19,10], Lea-Maria Löbel[10], Philipp Lutscher[80], Matthias Mader[72], Joan E. Madia[41], Natalia Malancu[67], Luis Maldonado[44], Helge-Johannes Marahrens[21], Nicole Martin[76], Paul Martinez[90], Jochen Mayerl[6], Oscar J. Mayorga[60], Patricia McManus[21], Kyle McWagner[49], Cecil Meeusen[27], Daniel Meierrieks[73], Jonathan Mellon[76], Friedolin Merhout[63], Samuel Merk[66], Daniel Meyer[62], Leticia Micheli[30], Jonathan Mijs[16,8], Cristóbal Moya[94], Marcel Neunhoeffer[37], Daniel Nüst[78], Olav Nygård[31], Fabian Ochsenfeld[34], Gunnar Otte[24], Anna Pechenkina[86], Christopher Prosser[46], Louis Raes[51], Kevin Ralston[91], Miguel Ramos[58], Arne Roets[12], Jonathan Rogers[39], Guido Ropers[37], Robin Samuel[74], Gregor Sand[59], Ariela Schachter[89], Merlin Schaeffer[63], David Schieferdecker[9], Elmar Schlueter[68], Regine Schmidt[57], Katja M. Schmidt[10], Alexander Schmidt-Catran[13], Claudia Schmiedeberg[32], Jürgen Schneider[82], Martijn Schoonvelde[54], Julia Schulte-Cloos[32], Sandy Schumann[55], Reinhard Schunck[84], Jürgen Schupp[10], Julian Seuring[57], Henning Silber[11], Willem Sleegers[51], Nico Sonnta[g75], Alexander Staudt[20], Nadia Steiber[83], Nils Steiner[24], Sebastian Sternberg[26], Dieter Stiers[27], Dragana Stojmenovska[56], Nora Storz[87], Erich Striessnig[83], Anne-Kathrin Stroppe[11], Janna Teltemann[71], Andrey Tibajev[31], Brian Tung[89], Giacomo Vagni[55], Jasper Van Assche[12,27], Meta van der Linden[8], Jolanda van der Noll[70], Arno Van Hootegem[27], Stefan Vogtenhuber[22], Bogdan Voicu[45,77], Fieke Wagemans[38], Nadja Wehl[93,57,72], Hannah Werner[27], Brenton M. Wiernik[81], Fabian Winter[35], Christof Wolf[11], Yuki Yamada[28], Nan Zhang[35], Conrad Ziller[64], Stefan Zins[23], Tomasz Żółtak[43]

Version: 26-Nov-2021

Summary Paragraph:

This study explores how analytical choices of researchers affect the reliability of scientific findings. Current lack-of-reliability discussions focus on systematic biases. We broaden the lens to include idiosyncratic decisions in data analysis that lead researchers to diverging results and conclusions. We coordinated and observed decisions among 73 research-teams as they independently tested the same hypothesis using the same data. Results show that in this typical secondary data research situation, the universe of pathways from data to results is so vast that each analysis was unique in some way. Teams reported divergent findings with contradictory substantive implications that could not be explained by differences in researchers' expertise, prior beliefs, and expectations. This calls for greater humility and clarity in presentation of scientific findings. Idiosyncratic variation may also be a cause for why many hypotheses remain highly contested, particularly in large-scale social and behavioral research.

Organized scientific knowledge production involves institutionalized checks such as editorial vetting, peer-review, and methodological standards to ensure that findings are independent of the characteristics or predispositions of any single researcher[1,2]. These procedures should generate inter-researcher reliability, offering consumers of scientific findings assurance that they are not arbitrary flukes and that other researchers would generate similar findings given the same data. Recent meta-science research challenges this assumption as many attempts to reproduce the findings of previous studies failed[3,4]. In response, scientists discuss various threats to the reliability of the scientific process.

These discussions tend to focus on a lack of reliability due to biases inherent in the production of science in practice. Pointing to both misaligned structural incentives and the cognitive tendencies of researchers[5–7], this bias-focused perspective argues that systematic distortions of the research process push the published literature away from truth and accurate observation. This then reduces the probability that a carefully executed replication will arrive at the same findings.

Here, we argue that the roots of reliability issues in science run even deeper systematically distorted research practices. We propose that to better understand why research is often non-replicable or lacking inter-researcher reliability we need to account for idiosyncratic variation inherent in the scientific process. Our main argument is that researcher variability can occur even under rigid adherence to the scientific method, high ethical standards and state-of-the-art approaches to maximizing reproducibility. As we report below, even well-meaning scientists freed from pressure to distort results and provided with identical data may not reliably converge in their findings because of the complexity and ambiguity inherent to the process of scientific analysis.

Variability in research outcomes

The scientific process confronts researchers with a multiplicity of seemingly minor, yet nontrivial, decision points. Each of these decision points may introduce variability in research outcomes. An important but underappreciated fact is that this even holds for what is often seen as the most objective step in the process: working with the data after it has come in. This problem may become particularly acute with large social and behavioral observational data. Researchers can take many different paths in wrangling, analyzing, presenting, and interpreting their data. Literally millions of different analytical pathways are possible for any given dataset, as the number of choices grows exponentially with the number of cases and variables included[8–10].

The bias-focused perspective on research reliability implicitly assumes that reducing incentives to generate surprising and sleek results would allow researchers to take a path that will lead to valid conclusions. This hope may have been too optimistic. While removing these barriers may prevent researchers from systematically taking invalid analytical paths[8–11], this alone does not guarantee researchers will converge on paths leading to valid outcomes. They could also disperse in different directions in the 'garden of forking paths' (following the term popularized by Gelman and Loken[8]). We just do not know much about the reliability of the data-analytic process.

A first approach to assessing and explaining data-analytical variation is to consider the individuals doing the data analysis: do their decisions vary based on how well-versed they are in applying relevant methods or their preexisting beliefs about what they will find? The *competency hypothesis* posits that researchers may make different analytical choices as a result of varying levels of statistical and subject expertise which leads to different judgments as to what constitutes the 'ideal' analysis in a given research situation. The *confirmation bias hypothesis* holds that researchers may make reliably different analytical choices as a result of differences in preexisting beliefs and attitudes, which may lead to justification of analytical approaches favoring certain outcomes post hoc. However, many other covert influences, large and small, may also lead to unreliable - and thus unexplainable, idiosyncratic variation in analytical decision pathways[10]. Crucially, even when distinct pathways appear equally reasonable to outsiders, seemingly minor variations between them may lead to widely varying outcomes.

There is growing awareness of the dependence of findings on statistical modeling decisions and the importance of analytical robustness[9,11,13,14]. But only recently scientists began to assess whether researcher variability affects scientific outcomes in realistic settings, sometimes employing 'many analysts, one dataset' approaches. For instance, when 29 researchers tested if soccer referees were biased toward darker skin players using the same data, 29 unique model specifications were reported with empirical results ranging from modestly negative to strongly positive[15]. Most of these many analyst studies were small in scale, preventing the possibility for multivariate meta-analysis of their results[16–18]. Some of these studies focused on the development of narrow, field-specific methods of analysis[17,19,20]. The aforementioned 'soccer study' was conducted in a research environment deliberately obtuse from involved researchers' substantive knowledge and foci. This strategy alleviates potentially biasing incentives or expectations but it also creates a less ecologically valid research setting. Nonetheless, these studies provide evidence that researchers are likely to come to a variety of results when provided with the same data and hypothesis even after incentives are removed; however, they provide little information as to why this is the case or how much their unbiased decisions shape the outcomes.

With most of these studies employing less than 30 researcher teams, reliable analysis of variance was either not possible or not even considered by the principal investigators. To circumvent this, the PIs of one study simulated researcher decisions to create a theoretical set of all outcomes that would have resulted had they had enough researchers to make all possible 'plausible' choices, known as a *metaverse* analysis. They then meta-analyzed how those simulated decisions impacted the simulated outcomes and used this as analogous evidence for how decisions in the workflow impacted results[21]. The drawback of a simulation approach is that it is constructed based on a single data analysis pipeline from one research team and not likely reflective of the more complex reality of different research processes carried out by different teams, which involves minute and seemingly trivial decisions (e.g., the choice of statistical analysis software).

A recent study by Botvinik-Nezer et al.[20] upped the ante by getting together an impressive 65 teams. This would appear to be enough to perform a multivariate analysis of outcome variance, and they were able to show two discrete factors that varied across teams could explain as much as 4% of the variance in outcomes: the software package used and estimated smoothness (a factor in the statistical analysis of functional magnetic resonance imaging, which was the data type used in the study). This finding suggested that if we were to observe every step of each independent research teams' workflows we might be able to explain even more of the outcome variance and arrive at a deeper understanding of variation, perhaps to the point where we could identify key moderator variables in the data that can explain why results go in a certain direction or have a certain size. Our study was designed to produce such knowledge.

We expected that a large controlled study of many analyst teams, allowing for close observation of the workflows of each team through surveys, pre-defined tasks and careful qualitative analysis of their code, would provide the necessary information to explain why highly skilled and accuracy-motivated researchers arrive at widely differing results when analyzing the same data. We assumed that certain consequential and identifiable decisions taken in each team's workflow such as their measurement strategy for the dependent and independent variables, selection of variables to include or omit in their models, estimation strategy and sample subsetting would explain how and why teams came to different results and conclusions. This would then allow us to examine how these decisions interacted to produce variation in the results of different researchers when they tested the same social science hypothesis with the same data. Meanwhile, we also assumed there would be *minor* analytical decisions (e.g., robust clustering, listwise deletion, model specification) that would matter but only generate some noise at the margins. We expected we could adjudicate between minor and major decisions, in addition to controlling for potential sources of variance in findings by observing all decisions of each team as well as their methods expertise and preexisting beliefs about the hypothesis allow us to test, and control for, both the competency and confirmation bias explanations for variation in researchers' data-analytical decision-making. Using all decisions, competencies and beliefs as variables we expected to explain outcome variance using multivariate regression approaches.

Design

The principal investigators (PIs) coordinated a group of 162 researchers in 73 teams who were given a simple instruction: to test, independently from each other, a hypothesis that has been at the center of an "extensive body of scholarship"[22]. Namely, whether immigration reduces support for social policies among the public. The hypothesis a typical example of social science studies that often test complex and broad hypotheses that leave room for interpretation regarding the central concepts or quantities of interest[23,24]. For example in a classic study, economists Alberto Alesina and Edward Glaeser hypothesized that differences in North American and European social security systems trace back to immigration-generated ethnic diversity[25,26]. New waves of immigration subsequently led Alesina and hundreds of other scholars to apply this hypothesis to the potential retrenchment of these systems within Western Europe and across the globe. In short, the hypothesis given to participating teams was chosen because the PIs considered it influential and representative of the practices in contemporary social research, for example in sociology, economics, political science and geography[27–32].

To increase the ecological validity of the project, the PIs provided teams with survey data from the *International Social Survey Programme* which had recently also been used in a frequently cited study that investigated the same hypothesis.[22] The ISSP is a long-running, high-quality, multi-country survey that is widely used in the social sciences. It includes a six-question module on different social policies and measures of stock and flow of immigrants by country and year that were analyzed in the original study (see Communications in Supplementary Materials for sampling details). To remove potentially biasing incentives, all participating researchers were ensured co-authorship on the final paper regardless of their results. Before running their test models, participants were given preparatory tasks to familiarize them with the topic and asked to develop and pre-submit a research design to the PIs. Moreover, the researchers participated in surveys before and after the project to capture predispositions and perceptions of the research process.

Participating teams were instructed to report standardized marginal effect estimates, yet were autonomous to decide what models to run and report. The teams submitted a total of 1,261 models, often following the template of the predecessor study which investigated this question with the same data and analyzed different policy attitude outcomes independently and using multiple model specifications.[22] Participants also submitted substantive conclusions on whether the data supported or rejected the hypothesis. The analysis code was checked and then anonymized for public sharing by the PIs (Figs. S1,S2 and Tables S1,S2). Many teams submitted 12 models testing two different immigration measures predicting opinions about the six social policies (model N ranged from 1 to 124 models per team; mean = 17.3).

Results

Fig. 1 visualizes the substantial variation of results reported by 73 researcher teams who analyzed the same data. Results are diffuse. Little more than half the reported estimates were statistically not significantly different from zero at 95% CI, while a quarter were significantly different and negative, and 16.9 percent were significant and positive.

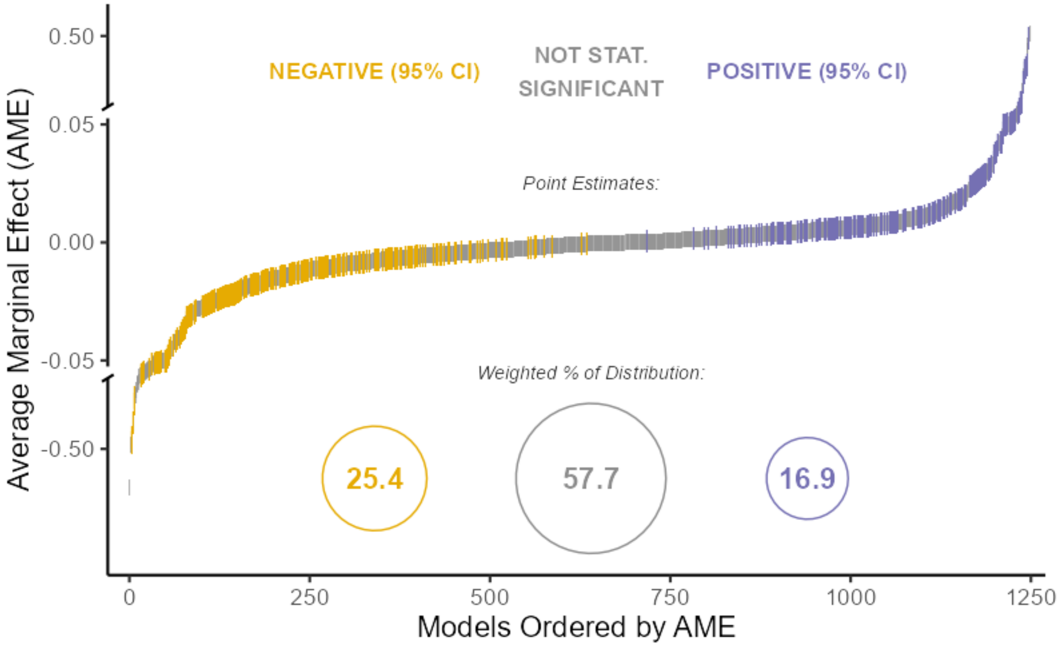

Fig. 1 Broad variation in findings from 73 teams testing the same hypothesis with the same data. The distribution of estimated average marginal effects (AME) across all converged models (N = 1,253) includes results that are negative (yellow, and in the direction predicted by the given hypothesis the teams were testing), not different from zero (grey) or positive (blue), using a 95% confidence interval. AME are XY-standardized. Y-axis contains two breaks at +/-0.05. Numbers inside circles represent the percentage of the distribution of each outcome inversely weighted by the number of models per team (see Interactive Results).

We observe the same pattern when we use the teams' subjective conclusions rather than their statistical results. Overall 13.5% (12 out of 89) of the team conclusions were that the hypothesis was *not testable* given these data, 60.7% (54 out of 89) concluded the hypothesis should be *rejected* and 28.5% (23 out of 89) concluded the hypothesis was *supported* (see Figs. S5,S9,S10). Note that 16 teams reported two differing conclusions based on their interpretation of different model specifications causing the N to jump from 73 teams to 89 team-conclusions.

We find that competencies and potential confirmation biases do not explain the broad variation in outcomes: researcher characteristics show no significant association with statistical results or even substantive conclusions (Fig. 2). Hence, it is not consistent with the data that outcome variability simply reflects a lack of knowledge among some participants or individual preferences for particular results.

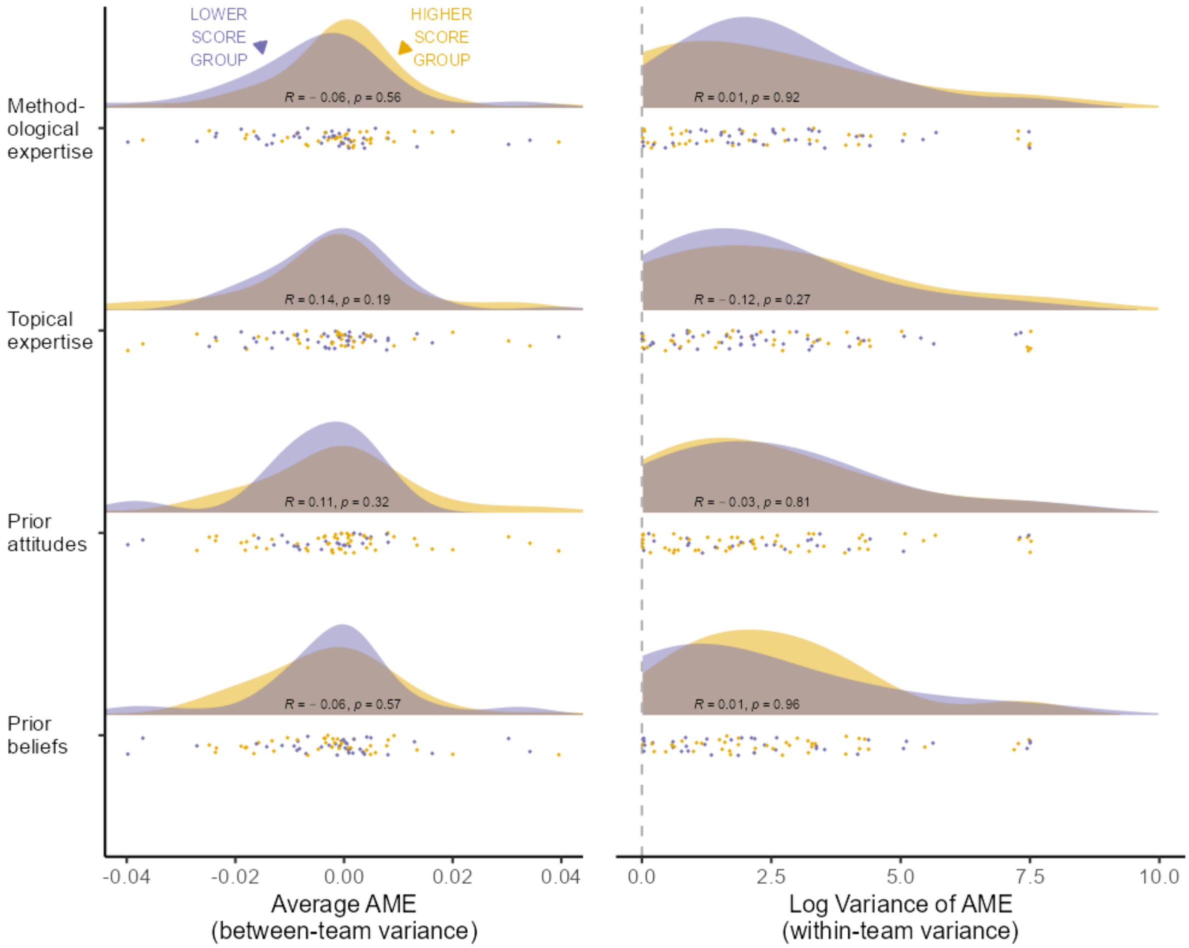

Fig. 2 Researcher characteristics do not explain outcome variance between teams or within teams. The distribution of team average of AMEs (left panel) and within-team variance in AMEs (right panel) across researchers grouped according to mean-splits ("LOWER" and "HIGHER") on methodological and topic expertise (potential *competencies bias*), and on prior attitudes toward immigration and beliefs about whether the hypothesis is true (potential *confirmation bias*). Log variance shifted so that minimum log value equals zero. Teams submitting only one model assigned a variance of zero. Pearson correlations along with a *p*-value ("R") are calculated using continuous scores of each researcher characteristic variable.

In the next step we attempted to locate the sources of outcome variability in the research process. We conducted an in-depth examination of all 1,261 models, which identified 166 research design decisions that were taken by at least one team. "Decision" here means any component of a model, for example measurement strategy, estimator, hierarchical structure, choice of independent variables and potential subsetting of the data (see Table S12). We found 107 identified decision points that were taken in more than two teams' workflows. Strikingly, the varying presence of these 107 common decisions in a dissimilarity matrix revealed that no two models were an identical combination (Table S4).

In principle, variation in outcomes must reflect prior decisions of the researchers. Yet, Fig. 3 shows that the 107 identified decision points explain very little of the variation. The major components of the identified researcher decisions explain less than a quarter of the variation in four measures of research outcomes. Most variance also remains unexplained after accounting for researcher characteristics or assignment to a small experiment (not reported in this study, see box "Assigned Conditions" in Fig. 3). Looking at total variance in the numerical results (top bar), identified components of the research design explain 2.6% (green segment) and researcher characteristics only account for a maximum of 1.2% of the variance. In other words, 95.2% of the total variance in results is left unexplained, suggesting that massive variation in reported results originated from idiosyncratic decisions in the data analysis process

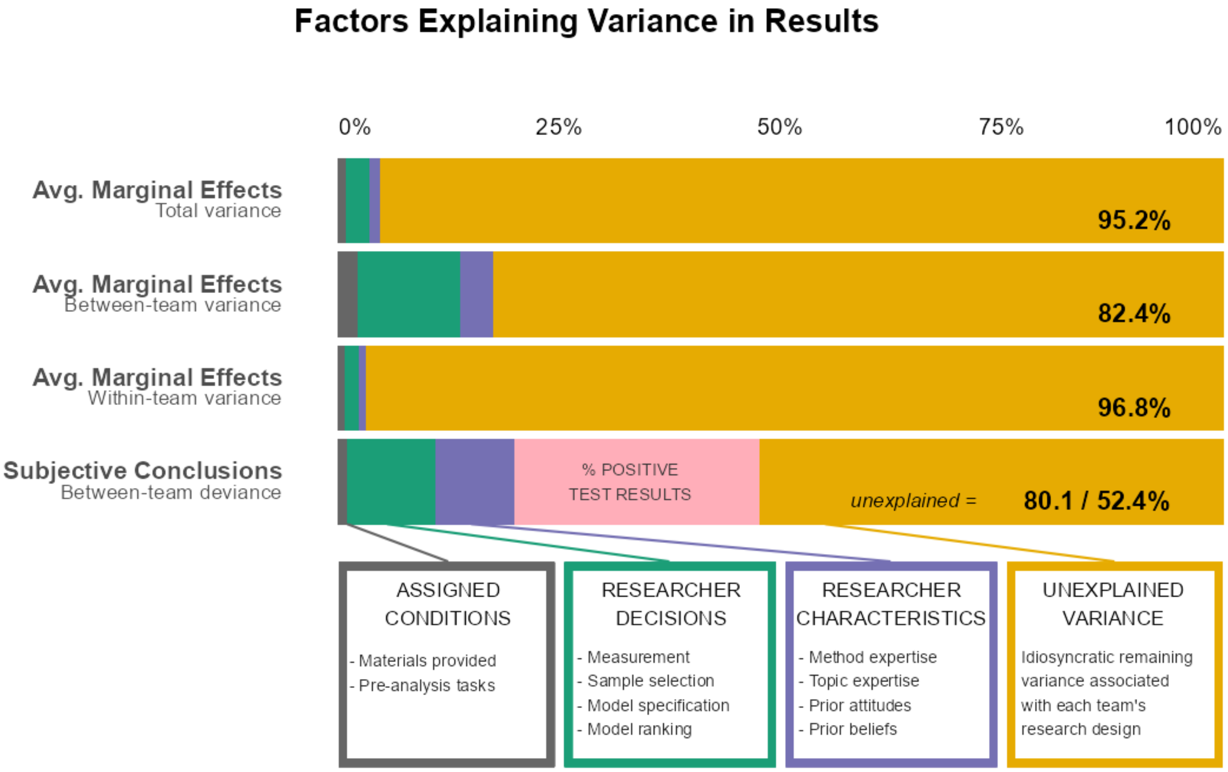

Fig. 3 Variance in statistical results and substantive conclusions between and within teams is mostly unexplained by conditions, research design and researcher characteristics. Decomposition of variance from generalized linear, multilevel regression models. Numerical outcomes (AMEs) (top three bars), and explained deviance from multinomial logistic regressions using the substantive conclusions about the target hypothesis as the outcome (bottom bar) submitted by the research teams. We used informed stepwise addition and removal of predictors to identify which specifications could explain the most numeric variance (Table S6) and others that could explain the most subjective conclusion deviance (Table S7) while sacrificing the least degrees of freedom and maintaining the highest level of model fit based on log-likelihood and various information criteria. We also tested every possible combination as a sensitivity check. Assigned conditions were the division of participants into two different task groups and two different deliberation groups during the preparatory phase. Identified researcher decisions are the 107 common decisions taken in data preparation and statistical modeling across teams and their models. Researcher characteristics were identified through a survey of participants and multi-item scaling using factor analysis (Fig S3). Many more other details throughout the Supplementary Materials.

The share of explained variance is somewhat higher when looking at between-team results (second-from-top bar) but still 82.4% remained unexplained. Variance remains mostly unexplained when moving away from the numerical results and when we look at the substantive conclusions that the researchers have drawn (bottom bar, 80.1% unexplained). It is noteworthy that even the percentage of test results per team that statistically support their conclusions, at least from a naïve 95% CI perspective, explain only about a third of the deviance in conclusions (salmon-colored segment, bottom bar), which points at variation in how different researchers interpret the same set of numerical results. Overall, the complexity of the data-analytic process leads to variation that cannot be easily explained even with a close look at researcher characteristics and researcher decisions.

We confirmed the robustness of our results by automatically assessing every possible combination of model specifications. We ran separate regressions by dependent variable and a variance function regression to check if specifications impacted the variability of results within teams and to correct for potential heteroscedasticity (Tables S4,S9-S11). All results support the conclusion that nearly all of the variance in research outcomes is from idiosyncratic researcher variability - unique sets of analytical decisions taken for each model in each team.

To gauge whether decisions explaining 2.6% of variance was a surprising finding or not, we ran a multiverse analysis. In this we found that using only 23 simulated decisions (dependent variable, three different test variable measurement strategies, an interaction of the previous two items, four different estimators, three different sample wave subsets, three different sample countries subsets, inclusion of up to three country-level 'control' variables and two different ways of introducing variance components) we could explain just over 16% of the variance in simulated numerical outcomes of 2,304 models (Table S8). This leaves a wide gap that we attribute to idiosyncratic researcher variability.

Summary

Results from our tightly controlled research design in a large-scale crowdsourced research effort involving 73 teams, demonstrate that offering the same data and hypothesis led to substantial variation in statistical estimates and substantive conclusions. Our finding of outcome variability echoes those of recent studies involving many researchers undertaken across scientific disciplines. It differs from these previous studies because it attempted to catalog every decision in the research process among each team and use those decisions and predictive modeling to explain why there is so much outcome variability. Every model was a unique combination of the 107 identified decisions taken across teams and across their models. In spite of this highly granular decomposition of the analytical process we could only explain less than 2.6% of the variance in numerical outcomes. We also tested if expertise, beliefs and attitudes observed among the teams biased results and they explained little. Even highly skilled scientists motivated to come to accurate results varied tremendously in what they found based on the same provided data and hypothesis. Our conclusion is that the totality of research decisions in the research process remain undisclosed in the standard presentation and consumption of scientific results. We have tapped into a hidden universe of idiosyncratic researcher variability.

Only many-analyst studies that observe researchers in their ecological environment can truly reveal this hidden universe. Simulations simply cannot capture the complexities and actual decisions taken by any given research team. For instance, a multiverse analysis allows for the decomposition of possible outcomes by simulating decisions, usually all possible combinations of a set of plausible decisions. Our own multiverse analysis explains about 16% of the variance in AMEs using a subset of consequential decisions. But we have no reason to believe that researchers would consider all possible combinations of certain decisions as plausible when they attempt to create a model of the data-generating process. Hence, previously used methods have failed to appreciate the role of both smaller and larger decisions researchers make in the research process that may be hidden from conscious awareness but are consequential for its outcomes, including the actual judgment calls they need to make about what they consider plausible analytical choices.

Implications and limitations

If researchers are responsible for assessing and communicating uncertainty, they should address sources of error. Their task is to recover a signal while attenuating noise as much as possible. Attenuation requires understanding the noise itself. In line with work that has demonstrated the implications of noise in human judgements[33,34], our study raises awareness of noise in the research process and enhances our understanding of the nature and magnitude of idiosyncratic variation that results.

Researchers must make analytical decisions so minute that they often do not register as decisions but go unnoticed as non-deliberate actions within ostensibly standard operating procedures. When taken as a whole, we show these hundreds of decisions are far from trivial and effect outcomes beyond the typically expected parsing or software-induced variabilities[35,36]. Moving forward from discussions of the reproducibility crisis in science, our findings suggest reliability across researchers may remain low even when their accuracy motivation is high and biasing incentives are removed. Higher levels of methodological expertise, another frequently suggested remedy, did not reduce variance either. Hence, we are left to believe that noise is a fundamental feature of the scientific process that is not easily explained by typically observed researcher characteristics or analytical decisions.

We believe that serious acknowledgment of idiosyncratic variation in research findings has at least three implications for improving presentation and interpretation of empirical evidence. First, contemplating that results would vary greatly if any given study had been conducted by a different set of researchers, or even the same researchers at a different time, underscores the uncertainty inherent in scientific outcomes. Drawing conclusions based on seemingly objective quantitative procedures therefore calls for epistemic humility. Second, the findings remind us to carefully document all analytical decisions, because the most seemingly minute could drive results in different directions. Third, countering the risk of defeatist notions, this study helps us appreciate the knowledge accumulated in those scientific fields where scientists do repeatedly converge on expert consensus - such as climate science or predictions of special relativity.

At the same time, we see limitations. First, we do not know the generalizability of our ecological setting to different topics, disciplines or even datasets. A major segment of social science works with observational survey data and our results directly reflect this type of research. In experimental research, the data-generating model is often much clearer or simply involves less decisions. Moreover, in the social sciences there are no Newtonian laws or definite quantum statistical likelihoods to work with, suggesting our case might be less conservative than a similar study in the natural sciences. It remains to be seen whether these data are more or less prone than other data to idiosyncratic variance in scientific results. Second, although we hoped to offer deeper insights on the substantive hypothesis under observation, we did not obtain evidence that moves conclusions in any direction. The lessons combined with the fact that a substantial portion of participants considered the hypothesis not testable with these data, offer a potential explanation for why this is such a contested hypothesis in the social science literature[22,29,37]. It brings forth a final general call for more attention to conceptual, causal and theoretical clarity, and gathering new data when results no longer appear to move a substantive area forward[23,24]. Our study adds to the urgency of this call. For us to reap epistemic benefits from the present move towards open research practices, we must complement transparency with theoretical clarity. Future *many-analyst* studies should consider this before investing such large amounts of resources in a new study.

We hope that others will embrace the noisy component of scientific knowledge, not as disheartening but as simply helping us to understand what we do and do not know - which may be a basic goal of science to begin with. Our findings demonstrate that we need to do more work to understand the idiosyncrasies across researchers in the data analysis process and the differences in their subjective conclusions. Rather than signifying a ‘crisis’, we submit that such work will bring ‘new opportunities and challenges’ for scientific advancement[38–40]. And that novel research and theories often arise out of contradictions[41].

We close by noting that the conclusions of this study were themselves derived from myriad seemingly minor analytical decisions, just like those we observed among the teams. We therefore encourage readers to scrutinize our analytical process and, to that end, provide robustness checks by taking advantage of the Supplementary Materials’ many additional figures and tables, complete reproduction files and a web-based interactive app that allow readers to easily explore the data themselves.

References and Notes

1. Solomon, M. *Social Empiricism*. (MIT press, 2007).
2. Oreskes, N. *Why Trust Science?* (Princeton University Press, 2019).
3. Camerer, C. F. *et al.* Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* **2**, 637–644 (2018).
4. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, (2015).
5. Ritchie, S. *Science Fictions: How Fraud, Bias, Negligence, and Hype Undermine the Search for Truth*. (Metropolitan Books, 2020).
6. Sørensen, A. B. The Structural basis of Social Inequality. *American Journal of Sociology* **101**, 1333–1365 (1996).
7. Frey, B. S. Publishing as Prostitution? – Choosing Between One's Own Ideas and Academic Success. *Public Choice* **116**, 205–223 (2003).
8. Gelman, A. & Loken, E. The Statistical Crisis in Science. *American Scientist* **102**, 460 (2014).
9. Orben, A. & Przybylski, A. K. The association between adolescent well-being and digital technology use. *Nature Human Behaviour* 173–189 (2019) doi:10.1038/s41562-018-0506-1.
10. Del Giudice, M. & Gangestad, S. A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions. *Advances in Methods and Practices in Psychological Science* (2020).
11. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* **22**, 1359–1366 (2011).
12. Analytical Methods Committee. Dark uncertainty. *Anal Methods* **4**, 2609–2612 (2012).
13. Schimmack, U. A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology/Psychologie canadienne* **61**, 364–376 (2020).
14. Freese, J. & Peterson, D. The Emergence of Statistical Objectivity: Changing Ideas of Epistemic Vice and Virtue in Science. *Sociological Theory* **36**, 289–313 (2018).
15. Silberzahn, R. *et al.* Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science* **1**, 337–356 (2018).

16. Landy, J. F. *et al.* Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin* (2020).

17. Dutilh, G. *et al.* The Quality of Response Time Data Inference: A Blinded, Collaborative Assessment of the Validity of Cognitive Models. *Psychon Bull Rev* **26**, 1051–1069 (2019).

18. Huntington-Klein, N. *et al.* The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry* **n/a**,.

19. 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. *Journal of Psychosomatic Research* **137**, 110211 (2020).

20. Botvinik-Nezer, R. *et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020).

21. Schweinsberg, M. *et al.* Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes* (2021).

22. Brady, D. & Finnigan, R. Does Immigration Undermine Public Support for Social Policy? *American Sociological Review* **79**, 17–42 (2014).

23. Auspurg, K. & Brüderl, J. Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the "Many Analysts, One Data Set" Project. *Socius* **86**, 532–565 (2021).

24. Lundberg, I., Johnson, R. & Stewart, B. M. What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *Am Sociol Rev* **86**, 532–565 (2021).

25. Alesina, A. & Glaeser, E. L. Why Are Welfare States in the US and Europe So Different? What Do We Learn? *Horizons stratégiques* **2**, 51–61 (2006).

26. Alesina, A. & Glaeser, E. *Fighting poverty in the US and Europe: A world of difference*. (Oxford University Press, 2004).

27. Eger, M. A. Even in Sweden: The Effect of Immigration on Support for Welfare State Spending. *European Sociological Review* **26**, 203–217 (2010).

28. Garand, J. C., Xu, P. & Davis, B. C. Immigration Attitudes and Support for the Welfare State in the American Mass Public. *American Journal of Political Science* **61**, 146–162 (2017).

29. Alesina, A., Murard, E. & Rapoport, H. Immigration and preferences for redistribution in Europe. *Journal of Economic Geography* (2021) doi:10.1093/jeg/lbab002.

30. Burgoon, B. Immigration, Integration, and Support for Redistribution in Europe. *World Politics* **66**, 365–405 (2014).
31. Alt, J. & Iversen, T. Inequality, Labor Market Segmentation, and Preferences for Redistribution. *American Journal of Political Science* **61**, 21–36 (2017).
32. Muñoz, J. & Pardos-Prado, S. Immigration and Support for Social Policy: An Experimental Comparison of Universal and Means-Tested Programs. *Political Science Research and Methods* **7**, 717–735 (2019).
33. Black, F. Noise. *The Journal of Finance* **41**, 528–543 (1986).
34. Kahneman, D., Sibony, O. & Sunstein, C. R. *Noise. A Flaw in Human Judgement*. (Little, Brown Spark, 2021).
35. Uhlmann, E. L. *et al.* Scientific Utopia III: Crowdsourcing Science. *Perspect Psychol Sci* **14**, 711–733 (2019).
36. McCoach, D. B. *et al.* Does the Package Matter? A Comparison of Five Common Multilevel Modeling Software Packages: *Journal of Educational and Behavioral Statistics* **43**, 594–627 (2018).
37. Eger, M. A. & Breznau, N. Immigration and the Welfare State: A Cross-Regional Analysis of European Welfare Attitudes. *International Journal of Comparative Sociology* **58**, 440–463 (2017).
38. Franco, A., Malhotra, N. & Simonovits, G. Publication bias in the social sciences: Unlocking the file drawer. *Science* **345**, 1502–1505 (2014).
39. Silberzahn, R. & Uhlmann, E. L. Crowdsourced Research: Many Hands make Light Work. *Nature* **526**, 189–191 (2015).
40. Fanelli, D. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *PNAS* **115**, 2628–2631 (2018).
41. Yanai, I. & Lercher, M. Novel predictions arise from contradictions. *Genome Biol* **22**, 153 (2021).

Acknowledgments: We are grateful to the Mannheim Centre for European Social Research at the University of Mannheim, Germany for providing extra institutional support for this large-scale project, and Joachim Gassen at the Humboldt University, Berlin, Germany for inspiration and feedback for our interactive app. Additionally helpful were comments provided at the Stanford METRICS International Forum, the LMU Sociology Research Colloquium, the TU Chemnitz 'Open Science' Conference, and in feedback from Steve Velsko, of the Lawrence Livermore National Lab, Livermore, California.

Funding:

Wikiversity Freies Wissen Fellowship 2019/20 (NB)

Author contributions:

Conceptualization: NB, EMR, AW

Data collection: NB, EMR, AW

Primary data analysis: NB

Data preparation, analysis assistance: NB, HHVN

Project administration: NB

Writing: EMR, AW, NB

GitHub Repository: NB, HHVN, DNÜST

Interactive app: NB, HHVN

Conflict of interests: Authors declare that they have no competing interests.

Data and materials availability: All code and data supporting the findings of this study have been deposited in GitHub at https://github.com/nbreznau/CRI.

Supplementary Materials

Materials and Methods

Figs. S1 to S12

Tables S1 to S12

Supplementary References (*42–45*)

42. H. Wickham, ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York, 2016; https://ggplot2.tidyverse.org).

43. M. Allen, D. Poggiali, K. Whitaker, T. R. Marshall, J. van Langen, R. A. Kievit, Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res*. 4, 63 (2021).

44. Y. Rosseel, lavaan: An R Package for Structural Equation Modeling. *J. Stat. Softw*. 48, 1–36 (2012).

45. N. Breznau, E. M. Rinke, A. Wuttke, Pre-Analysis Plan: Measuring Routine Researcher Variability in Macro-Comparative Secondary Data Analyses (2018), (available at https://osf.io/weu2v/).

Interactive Results (web-based):

https://nate-breznau.shinyapps.io/shiny/