University of St. Thomas, Minnesota

UST Research Online

Statistics Faculty Publications | Statistics

2021

Same Data, Different Conclusions: Radical Dispersion in Empirical Results When Independent Analysts Operationalize and Test the Same Hypothesis

Martin Schweinsberg

Michael Feldman

Nicola Staub

Olmo R. van den Akker

Robbie C.M. van Aert

See next page for additional authors

Follow this and additional works at: https://ir.stthomas.edu/cas_stats_pub

Recommended Citation

Schweinsberg, Martin; Feldman, Michael; Staub, Nicola; van den Akker, Olmo R.; van Aert, Robbie C.M.; van Assen, Marcel A.L.M.; Liu, Yang; Althoff, Tim; Heer, Jeffrey; and McNamara, Amelia, "Same Data, Different Conclusions: Radical Dispersion in Empirical Results When Independent Analysts Operationalize and Test the Same Hypothesis" (2021). *Statistics Faculty Publications*. 17.
https://ir.stthomas.edu/cas_stats_pub/17

This Article is brought to you for free and open access by the Statistics at UST Research Online. It has been accepted for inclusion in Statistics Faculty Publications by an authorized administrator of UST Research Online. For more information, please contact asle4660@stthomas.edu.

Author

Martin Schweinsberg, Michael Feldman, Nicola Staub, Olmo R. van den Akker, Robbie C.M. van Aert, Marcel A.L.M. van Assen, Yang Liu, Tim Althoff, Jeffrey Heer, and Amelia McNamara

This article is available at UST Research Online: https://ir.stthomas.edu/cas_stats_pub/17

Organizational Behavior and Human Decision Processes 165 (2021) 228–249

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis☆

Martin Schweinsberg [a,1,*], Michael Feldman [b,1,*], Nicola Staub [b,1], Olmo R. van den Akker [c], Robbie C.M. van Aert [c], Marcel A.L.M. van Assen [d], Yang Liu [e], Tim Althoff [e], Jeffrey Heer [e], Alex Kale [e], Zainab Mohamed [f], Hashem Amireh [g], Vaishali Venkatesh Prasad [a], Abraham Bernstein [b,*], Emily Robinson, Kaisa Snellman [h], S. Amy Sommer [i], Sarah M.G. Otner [j], David Robinson, Nikhil Madan [k], Raphael Silberzahn [l], Pavel Goldstein [m], Warren Tierney [n], Toshio Murase [o], Benjamin Mandl [p], Domenico Viganola [p], Carolin Strobl [b], Catherine B. C. Schaumans [q], Stijn Kelchtermans [r], Chan Naseeb [s], S. Mason Garrison [t], Tal Yarkoni [u], C. S. Richard Chan [v], Prestone Adie [w], Paulius Alaburda, Casper Albers [x], Sara Alspaugh [y], Jeff Alstott [z], Andrew A. Nelson [aa], Eduardo Ariño de la Rubia [ab], Adbi Arzi [h], Štěpán Bahník [ac], Jason Baik, Laura Winther Balling [ad], Sachin Banker [ae], David AA Baranger [af], Dale J. Barr [ag], Brenda Barros-Rivera [ah], Matt Bauer [ai], Enuh Blaise [aj], Lisa Boelen [ak], Katerina Bohle Carbonell [al], Robert A. Briers [am], Oliver Burkhard, Miguel-Angel Canela [an], Laura Castrillo, Timothy Catlett, Olivia Chen, Michael Clark [ao], Brent Cohn, Alex Coppock [ap], Natàlia Cugueró-Escofet [aq], Paul G. Curran [ar], Wilson Cyrus-Lai [h], David Dai [as], Giulio Valentino Dalla Riva [at], Henrik Danielsson [au], Rosaria de F.S.M. Russo [av], Niko de Silva [a], Curdin Derungs [aw], Frank Dondelinger [ax], Carolina Duarte de Souza [ay], B. Tyson Dube, Marina Dubova [az], Ben Mark Dunn [ag], Peter Adriaan Edelsbrunner [ba], Sara Finley [bb], Nick Fox [bc], Timo Gnambs [bd], Yuanyuan Gong [be], Erin Grand, Brandon Greenawalt [bf], Dan Han, Paul H.P. Hanel [bg], Antony B. Hong [h], David Hood, Justin Hsueh, Lilian Huang [bh], Kent N. Hui [bi], Keith A. Hultman [bj], Azka Javaid [bk], Lily Ji Jiang [bl], Jonathan Jong [bm], Jash Kamdar, David Kane [bn], Gregor Kappler [bo], Erikson Kaszubowski [ay], Christopher M. Kavanagh, Madian Khabsa, Bennett Kleinberg [bp], Jens Kouros, Heather Krause [bq], Angelos-Miltiadis Krypotos [br], Dejan Lavbič [do], Rui Ling Lee [bs], Timothy Leffel [bh], Wei Yang Lim [bt], Silvia Liverani [bu], Bianca Loh [h], Dorte Lønsmann [bv], Jia Wei Low [bw], Alton Lu [e], Kyle MacDonald [bx], Christopher R. Madan [by], Lasse Hjorth Madsen [bz], Christina Maimone [al], Alexandra Mangold, Adrienne Marshall [ca], Helena Ester Matskewich [e], Kimia Mavon [bn], Katherine L. McLain [a], Amelia A. McNamara [cc], Mhairi McNeill, Ulf Mertens [cd], David Miller [al], Ben Moore [ce], Andrew Moore, Eric Nantz [cf], Ziauddin Nasrullah [a], Valentina Nejkovic [cg], Colleen S Nell [ch], Andrew Arthur Nelson [aa], Gustav Nilsonne [ci], Rory Nolan [cj], Christopher E. O'Brien, Patrick O'Neill [ck], Kieran O'Shea [ag], Toto Olita [cl], Jahna Otterbacher [cm], Diana Palsetia [al], Bianca Pereira, Ivan Pozdniakov [cn], John Protzko [co], Jean-Nicolas Reyt [cp], Travis Riddle [cq], Amal (Akmal) Ridhwan Omar Ali [cr], Ivan Ropovik [cs], Joshua M. Rosenberg [ct], Stephane Rothen, Michael Schulte-Mecklenbeck [cu], Nirek Sharma [cv], Gordon Shotwell [cw], Martin Skarzynski, William Stedden, Victoria Stodden [cx], Martin A. Stoffel [cy], Scott Stoltzman [cz],

☆ This article is an invited submission. It is part of the special issue "Best Practices in Open Science," Edited by Don Moore and Stefan Thau.

* Corresponding authors.

https://doi.org/10.1016/j.obhdp.2021.02.003

Available online 17 June 2021

0749-5978/© 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Subashini Subbaiah [da], Rachael Tatman [db], Paul H. Thibodeau [dc], Sabina Tomkins [dd], Ana Valdivia [de], Gerrieke B. Druijff-van de Woestijne [df], Laura Viana [dg], Florence Villesèche [ad], W. Duncan Wadsworth [dh], Florian Wanders [di], Krista Watts, Jason D Wells [dj], Christopher E. Whelpley [dk], Andy Won, Lawrence Wu [y], Arthur Yip, Casey Youngflesh [dl], Ju-Chi Yu [dm], Arash Zandian [dn], Leilei Zhang, Chava Zibman, Eric Luis Uhlmann [n,1,*]

[a] *ESMT Berlin, Germany*
[b] *University of Zurich, Switzerland*
[c] *Tilburg University, Netherlands*
[d] *Tilburg University and Utrecht University, Netherlands*
[e] *University of Washington, United States*
[f] *ESMT Berlin and Indiana University, Germany*
[g] *ESMT Berlin and Humboldt University Berlin, Germany*
[h] *INSEAD, France*
[i] *Marshall School of Business, University of Southern California, United States*
[j] *Imperial College Business School, United Kingdom*
[k] *Indian School of Business, India*
[l] *University of Sussex Business School, United Kingdom*
[m] *School of Public Health, University of Haifa, Israel*
[n] *INSEAD, Singapore*
[o] *Waseda University, Japan*
[p] *Stockholm School of Economics, Sweden*
[q] *Independent researcher*
[r] *KU Leuven, Belgium*
[s] *IBM, Germany*
[t] *Wake Forest University, United States*
[u] *University of Texas at Austin, United States*
[v] *Stony Brook University, United States*
[w] *University of Nairobi, Kenya*
[x] *University of Groningen, Netherlands*
[y] *University of California, Berkeley, United States*
[z] *Massachusetts Institute of Technology, United States*
[aa] *University of Kentucky, United States*
[ab] *California State University-Dominguez Hills, United States*
[ac] *The Prague College of Psychosocial Studies, Czech Republic*
[ad] *Copenhagen Business School, Denmark*
[ae] *University of Utah, United States*
[af] *University of Pittsburgh, United States*
[ag] *University of Glasgow, United Kingdom*
[ah] *Texas A&M University, United States*
[ai] *Illinois Institute of Technology, United States*
[aj] *Eskisehir Osmangazi University, Turkey*
[ak] *Imperial College London, United Kingdom*
[al] *Northwestern University, United States*
[am] *Edinburgh Napier University, United Kingdom*
[an] *University of Navarra, Spain*
[ao] *University of Michigan, United States*
[ap] *Yale University, United States*
[aq] *Universitat Oberta de Catalunya, Spain*
[ar] *Michigan State University, United States*
[as] *St. Michael's Hospital, University of Toronto, Canada*
[at] *Department of Mathematics and Statistics, University of Canterbury, New Zealand*
[au] *Linköping University, Sweden*
[av] *Universidade Nove de Julho, Brazil*
[aw] *Lucerne University of Applied Sciences and Arts, Switzerland*
[ax] *Lancaster University, United Kingdom*
[ay] *Universidade Federal de Santa Catarina, Brazil*
[az] *Indiana University, United States*
[ba] *ETH Zurich, Switzerland*
[bb] *Pacific Lutheran University, United States*
[bc] *Rutgers University, United States*
[bd] *Leibniz Institute for Educational Trajectories, Germany, & Johannes Kepler University Linz, Austria*
[be] *Okayama University, Japan*
[bf] *University of Notre Dame, United States*
[bg] *University of Bath, University of Essex, United Kingdom*
[bh] *University of Chicago, United States*
[bi] *School of Management, Xiamen University, China*
[bj] *Elmhurst College, United States*
[bk] *Columbia University Medical Center, United States*
[bl] *University of Washington & Indiana University, United States*
[bm] *University of Oxford & Coventry University, United Kingdom*
[bn] *Harvard University, United States*
[bo] *University of Vienna, Austria*
[bp] *University College London, United Kingdom*
[bq] *York University, United Kingdom*
[br] *Department of Clinical Psychology, Utrecht University, the Netherlands, & Group of Health Psychology, KU Leuven, Belgium*
[bs] *Nanyang Technological University, Singapore*
[bt] *University of Colorado, Colorado Springs, United States*

[bu] Queen Mary University of London, United Kingdom
[bv] University of Copenhagen, Denmark
[bw] Singapore Management University, Singapore
[bx] McD Tech Labs, United States
[by] School of Psychology, University of Nottingham, United Kingdom
[bz] Novo Nordisk, Denmark
[ca] University of Idaho, United States
[cc] University of St Thomas, United States
[cd] Heidelberg University, Germany
[ce] University of Edinburgh, United Kingdom
[cf] Eli Lilly, United States
[cg] University of Nis, Faculty of Electronic Engineering, Serbia
[ch] George Washington University, United States
[ci] Karolinska Institutet and Stockholm University, Sweden
[cj] University of Oxford, United Kingdom
[ck] University of Maryland, Baltimore County, United States
[cl] The University of Western Australia
[cm] Open University of Cyprus, Cyprus
[cn] National Research University, Higher School of Economics, Russia
[co] University of California, Santa Barbara, United States
[cp] McGill University, Canada
[cq] National Institutes of Health/National Institute of Mental Health, United States
[cr] The University of Sheffield, United Kingdom
[cs] Charles University, Faculty of Education, Institute for Research and Development of Education, Czech Republic & University of Presov, Faculty of Education, Slovakia
[ct] University of Tennessee, Knoxville, United States
[cu] University of Bern, Switzerland & Max Planck Institute for Human Development, Germany
[cv] Washington University in St. Louis, United States
[cw] Dalhousie University, Canada
[cx] University of Illinois at Urbana-Champaign, United States
[cy] Institute of Evolutionary Biology, University of Edinburgh, United Kingdom
[cz] Colorado State University, United States
[da] CSU, United States
[db] Rasa Technologies, United States
[dc] Oberlin College, United States
[dd] Stanford University, United States
[de] University of Granada, Spain
[df] Radboud University Nijmegen, Netherlands
[dg] University of Hawaii, United States
[dh] Microsoft & Rice University, United States
[di] University of Amsterdam, Netherlands
[dj] Dartmouth College, United States
[dk] College of Charleston, United States
[dl] Department of Ecology and Evolutionary Biology, University of California, Los Angeles, United States
[dm] The University of Texas at Dallas, School of Behavioral and Brain Sciences, United States
[dn] Division of Affinity Proteomics, Department of Protein Science, KTH Royal Institute of Technology & SciLifeLab, Sweden
[do] University of Ljubljana, Slovenia

ARTICLE INFO

Keywords:
Crowdsourcing data analysis
Scientific transparency
Research reliability
Scientific robustness
Researcher degrees of freedom
Analysis-contingent results

ABSTRACT

In this crowdsourced initiative, independent analysts used the same dataset to test two hypotheses regarding the effects of scientists' gender and professional status on verbosity during group meetings. Not only the analytic approach but also the operationalizations of key variables were left unconstrained and up to individual analysts. For instance, analysts could choose to operationalize status as job title, institutional ranking, citation counts, or some combination. To maximize transparency regarding the process by which analytic choices are made, the analysts used a platform we developed called DataExplained to justify both preferred and rejected analytic paths in real time. Analyses lacking sufficient detail, reproducible code, or with statistical errors were excluded, resulting in 29 analyses in the final sample. Researchers reported radically different analyses and dispersed empirical outcomes, in a number of cases obtaining significant effects in opposite directions for the same research question. A Boba multiverse analysis demonstrates that decisions about how to operationalize variables explain variability in outcomes above and beyond statistical choices (e.g., covariates). Subjective researcher decisions play a critical role in driving the reported empirical results, underscoring the need for open data, systematic robustness checks, and transparency regarding both analytic paths taken and not taken. Implications for organizations and leaders, whose decision making relies in part on scientific findings, consulting reports, and internal analyses by data scientists, are discussed.

[1] Author contributions. The first three and last author contributed equally to this project. MS coordinated the overall project. MS, MF, NS, AB, and EU conceptualized the project. MF, NS, & AB created the DataExplained platform. OvdA, RvA, and MvA carried out the quantitative analyses of the results of the overall project. YL, TA, JH and AK carried out the Boba multiverse analysis. ESR, KS, AS, SO, DR, NM, and RS constructed the dataset used in the project. ESR, KS, AS, and SO coordinated the pilot study. PG, WT, TM, BM, DV, HA, VP, ZM and CS provided further statistical expertise. MF and NS carried out the qualitative analyses of researcher justifications for their decisions. Authors 24 to 179 contributed hypotheses in the idea generation phase, analyzed data as part of the pilot, served as crowdsourced analysts for the primary project, and/or helped with project logistics. MS, MF, NS, OvdA, RvA, MvA, AB, & EU drafted the manuscript. All authors provided edits and feedback on the manuscript.

1. Introduction

In a typical scientific investigation, one researcher or a small team of researchers presents analytical results testing a particular set of research hypotheses. However, as many scholars have argued, there are often numerous defensible analytic specifications that could be used on the same data, raising the issue of whether variations in such specifications might produce qualitatively different outcomes (Bamberger, 2019; Cortina, Green, Keeler, & Vandenberg, 2017; Gelman, 2015; Gelman & Loken, 2014; Leamer, 1985; Patel, Burford, & Ioannidis, & 2015; Saylors & Trafimow, in press; Wicherts et al., 2016). This question generally goes unanswered, as most datasets from published articles are not available to peers (Aguinis & Solarino, in press; Alsheikh-Ali, Qureshi, Al-Mallah, & Ioannidis, 2011; Savage & Vickers, 2009; Vines et al., 2013; Wicherts, Borsboom, Kats, & Molenaar, 2006; Womack, 2015; Young & Horvath, 2015). However, simulations and case studies suggest that the exploitation of researcher degrees of freedom could easily lead to spurious findings (Simmons, Nelson, & Simonsohn, 2011), coding different research articles from the same topic area reveals as many analytic approaches as there are publications (Carp, 2012a, 2012b), and meta-scientific statistical techniques find evidence of publication bias, p-hacking, and otherwise unreliable results across various scientific literatures (e.g., O'Boyle, Banks, & Gonzalez-Mulé, 2017; O'Boyle, Banks, Carter, Walter, & Yuan, 2019; Williams, O'Boyle, & Yu, 2020). Multiverse analyses and specification curves, in which one analyst attempts many different approaches, suggest that some published conclusions only obtain empirical support in a small subset of specifications (Orben & Przybylski, 2019; Simonsohn, Simmons, & Nelson, 2020; Smerdon, Hu, McLennan, von Hippel, & Albrecht, 2020; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). Underscoring the pitfalls when published analyses of complex datasets focus on a single primary specification, two papers were recently published in the same surgical journal, analyzing the same large dataset and drawing opposite recommendations regarding Laparoscopic appendectomy techniques (Childers & Maggard-Gibbons, 2020).

In the crowdsourced approach to data analysis, numerous scientists independently analyze the same dataset to test the same hypothesis (Silberzahn & Uhlmann, 2015). If similar results are obtained by many analysts, scientists can speak with one voice on an issue. Alternatively, the estimated effect may be highly contingent on analysis strategies. If so, then subjectivity in applying statistical decisions and ambiguity in scientific results can be made transparent. The first crowdsourcing data analysis initiative examined potential racial bias in organizational settings, specifically whether soccer referees give more red cards to dark-skin toned players than to light-skin toned players (Silberzahn et al., 2018). The project coordinators collected a dataset with 146,028 referee-player dyads from four major soccer leagues and recruited 29 teams of analysts to test the hypothesis using whatever approach they felt was most appropriate. The outcome was striking: although approximately two-thirds of the teams obtained a significant effect in the expected direction, effect size estimates ranged from a nonsignificant tendency for light-skin toned players to receive more red cards to a strong tendency for dark-skin toned players to receive more red cards (0.89 to 2.93 in odds ratio units). Effect size estimates were similarly dispersed for expert analysts, and for analyses independently rated as high in quality, indicating variability in analytic outcomes was not due to a few poorly specified analytic approaches. This suggests that defensible, but subjective, analytic choices can lead to highly variable quantitative effect size estimates. The disturbing implication is that if only one team had obtained the dataset and presented their preferred analysis, the scientific conclusion drawn could have been anything from major racial disparities in red cards to equal outcomes.

Subsequent crowd initiatives have likewise revealed divergent results across independent scientific teams (Bastiaansen, Kunkels, & Blaauw, 2020; Botvinik-Nezer et al., 2020). Relying on fMRI data from 108 research participants who performed a version of a decision-making task involving risk, Botvinik-Nezer et al. (2020) recruited 70 research teams to test nine hypotheses (e.g., "Positive parametric effect of gains in the vmPFC"). Analysts were asked whether each hypothesis was supported overall (yes/no) in their analysis of the dataset. No two teams used the same approach, and only 1 of 9 hypotheses received support (i.e., a "yes" response) across the large majority of teams (Hypothesis 5, with 84.3% support). Three hypotheses were associated with nearly-uniform null results across analysts (94.3% non-significant findings), while for the remaining five hypotheses between 21.4% and 37.1% of teams reported statistically significant support. At the same time, *meta*-analysis revealed significant convergence across analysis teams in terms of the activated brain regions they each identified. In another recent crowd project, Bastiaansen et al. (2020) recruited 12 analysis teams with expertise in event sampling methods to analyze individual time-series data from a single clinical patient for the purposes of identifying treatment targets. A standard set of questionnaire items assessing depression and anxiety (e.g., "I felt a loss of interest or pleasure", 0 = not at all, 100 = as much as possible) was administered repeatedly to the same single patient over time. Participating researchers were asked "What symptom (s) would you advise the treating clinician to target subsequent treatment on, based on a person-centered (-specific) analysis of this particular patient's ESM data?" Analysts differed in their data preprocessing steps, statistical techniques, and software packages. The nature of identified target symptoms likewise varied widely (ranging between 0 and 16 targets), and no two teams made similar recommendations regarding symptoms to target for treatment.

The analysis-contingent results revealed via crowdsourcing represent a more fundamental challenge for scholarship across disciplines than *p*-hacking (selecting an analytic approach to achieve statistical significance; Banks, Rogelberg, Woznyj, Landis, & Rupp, 2016; Bedeian, Taylor, & Miller, 2010; O'Boyle et al., 2017; O'Boyle et al., 2019; Simmons et al., 2011) and peeking at the data and then testing for what look like significant relationships (Bosco, Aguinis, Field, Pierce, & Dalton, 2016; Gelman & Loken, 2014). The latter two threats to validity can be addressed by pre-registering the analytic strategy (Aguinis, Banks, Rogelberg, Cascio, in press; Banks et al., 2016, 2019; Van 't Veer & Giner-Sorolla, 2016; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012), or conducting a blinded analysis in which variables are temporarily changed (MacCoun & Perlmutter, 2015). In the latter approach variable labels might be switched (e.g., the Conscientiousness personality variable really refers to Agreeableness scores), or variable scores could be recoded (e.g., political conservatism is reverse coded such that high scores mean liberalism not conservatism). The key is that the reader does not know whether the observed relations among variables are consistent with her theoretical hypothesis or not. Under these circumstances, the researcher cannot consciously or unconsciously choose an analytic approach that produces statistically significant results in the hoped-for direction. In contrast, analysis-contingent results will still occur without perverse publication incentives because analysts, even if they act transparently and in good faith, are likely to use divergent approaches to answer the research question. Pre-registration or blinding data does not solve this because different investigators will preregister different analyses, and choose different approaches even with blinded data. Subjective choices and their consequences, often based on prior theoretical assumptions, may be an inextricable aspect of the scientific process.

2. The present research

There is good reason to believe that Silberzahn et al. (2018) in fact

underestimated the impact of researcher decisions on the results of a scientific investigation. Operationalizations of key theoretical variables were artificially restricted to red card decisions based on skin tone. Yet the conceptual research question ("Are referees biased by a player's race?") could have led to analyses involving yellow cards, stoppage time, offside calls, membership in specific ethnic groups, or indices of race and racial groups. Similarly, in Botvinik-Nezer et al.'s (2020) crowdsourced initiative using fMRI data, variability in results was due to methodological factors such as regressors, software packages, preprocessing steps, and demarcation of anatomical regions – not conceptualizations of the research question or theoretical constructs, which were narrowly defined. The experience sampling dataset used in Bastiaansen et al. (2020) was based on a set of standardized questionnaire items, with variability in results attributable to data preprocessing, statistical techniques, and software packages. Although different analysts clustered items differently, they did not employ fundamentally different approaches to conceptualizing and measuring variables like depression and anxiety. In contrast, in the present initiative crowdsourcing the analysis of a complex dataset on gender and professional status in group meetings, conceptualization and operationalization of key variables (e.g., social status) was left unconstrained and up to individual researchers. This approach is arguably closer to the ambiguity researchers typically confront when approaching a complicated dataset, and may lead to even greater heterogeneity of methods and results than seen previously.

The dataset for this project included over three million words and thousands of pieces of dialogue from an invitation-only online forum for scientific debates (see Supplement 1 for a detailed overview and https://osf.io/u9zs7/ for the dataset). Consider the simple and straightforward hypothesis that high status scientists tend to speak more during such group meetings. An analyst might choose to operationalize professional status using dataset variables such as citation counts, h-index, i10-index, job title, rankings of current university, rankings of doctoral institution, years since PhD, or some combination of the above. She might also decide to focus on professional status within a field, subfield, or among participants in an individual conversation, and use this to predict how actively the person participated in the meeting. Likewise, verbosity might be operationalized in different ways, among these number of words contributed, or number of comments made.

The overall project featured a pilot phase to generate and select hypotheses, and also carry out initial analyses testing these hypotheses (see Supplements 2 and 3 for detailed reports). To help generate and evaluate ideas, a crowd of scientists recruited online were provided with an overview of the dataset (variables and data structure) and asked to propose research hypotheses that might be tested with it. The crowd then voted on which ideas should be selected for systematic testing (Supplement 2). Subsequently, a small number of research teams (a subset of this crowd) used the dataset to test the final set of eleven hypotheses. As reported in Supplement 3, the quantitative results of these pilot analyses proved remarkably dispersed across teams, with little convergence in outcomes for any of the scientific predictions.

The primary study reported in the present manuscript reduced the number of hypotheses from eleven to two characterized by positive evaluations in the selection survey (Supplement 2) and divergent results in the pilot analyses (Supplement 3). We focused on two hypotheses from the pilot with especially dispersed outcomes across analysts in order to pursue our goal of understanding the sources of such variability. To this end, we asked analysts to use an online platform we developed called DataExplained to articulate the reasoning underlying each of their analytic decisions as they made them (further details on how the platform works are provided in the Methods section, in Feldman, 2018, Staub, 2017, and in Supplement 9). The stated reasons were then subjected to a qualitative analysis based on the General Inductive Approach (Thomas, 2006). DataExplained offers a novel form of scientific transparency, in that it documents analytic paths being taken and not taken in real time and provides this output in addition to the traditional research analytic outputs.

Both of the research ideas selected for crowdsourced testing were previously explored in the managerial and psychological literatures on gender, status, and group dynamics (Brescoll, 2011; Inzlicht & Ben-Zeev, 2000; Schmid Mast, 2001, 2002; Spencer, Logel, & Davies, 2016). Hypothesis 1 posits that "*A woman's tendency to participate actively in a conversation correlates positively with the number of females in the discussion.*" Hypothesis 2 predicts that "*Higher status participants are more verbose than are lower status participants*." Our project examined whether independent analysts would arrive at similar analyses and statistical results using the same dataset to address these questions.

In addition to recruiting a crowd of analysts to test Hypothesis 1 and 2, we carried out a complementary multiverse analysis using the Boba approach (Liu et al., 2020). A multiverse analysis evaluates all reasonable combinations between analytic choices (Simonsohn et al., 2020; Steegen et al., 2016), which in this case includes and expands beyond the paths taken by the crowd analysts. The Boba multiverse allows us to examine all "reasonable" paths implied by the juxtaposition of crowd submissions, quantitatively identify which choice points played the largest roles in effect size dispersion across analysts, and create visualizations illustrating some of the key steps in this garden of forking paths (Liu et al., 2020). To build the Boba multiverse, we took the key choice points faced by the analysts in the present project, and the major categories of approaches they used to dealing with them. Analysts had to choose the dataset variables they would use to capture the independent and dependent variables (e.g., whether to measure status with academic citations or job rank), determine their unit of analysis (e.g., commentators vs. conversations), decide what covariates to include, and which type of regression or other measure of association to use. In the Boba multiverse, we crossed as many choice as possible and was reasonable, and examined the implications for the final estimates for both Hypotheses 1 and 2.

3. Methods

3.1. Dataset

The dataset included 3,856,202 words of text in 7,975 comments from the online academic forum Edge (Lazer et al., 2009). As described by Edge's founders, its purpose is: "To arrive at the edge of the world's knowledge, seek out the most complex and sophisticated minds, put them in a room together, and have them ask each other the questions they are asking themselves" (http://edge.org). The group discussions spanned almost two decades (1996–2014) and included 728 contributors, 128 of them female. The dataset contained 150 variables related to the conversation, its contributors, or the textual level of the transcript (Supplement 1). New attributes not provided on the website were manually collected by browsing CVs, university or personal web-pages, Google Scholar pages, and professional networking websites, and added to the dataset.

An anonymized version of the dataset for the project is available at: https://osf.io/u9zs7/. The dataset is structured as follows: each row in the dataset presents one comment made by one contributor to one conversation. Each row contained variables for comment id, conversation id, and contributor id. Each comment contributed to only one conversation. A comment consisted of at least one character, and most comments consisted of several words and sentences. A new comment was created when a contributor wrote at least one character that was submitted to the forum. A conversation started when a contributor wrote a new comment that did not respond to a previous comment.

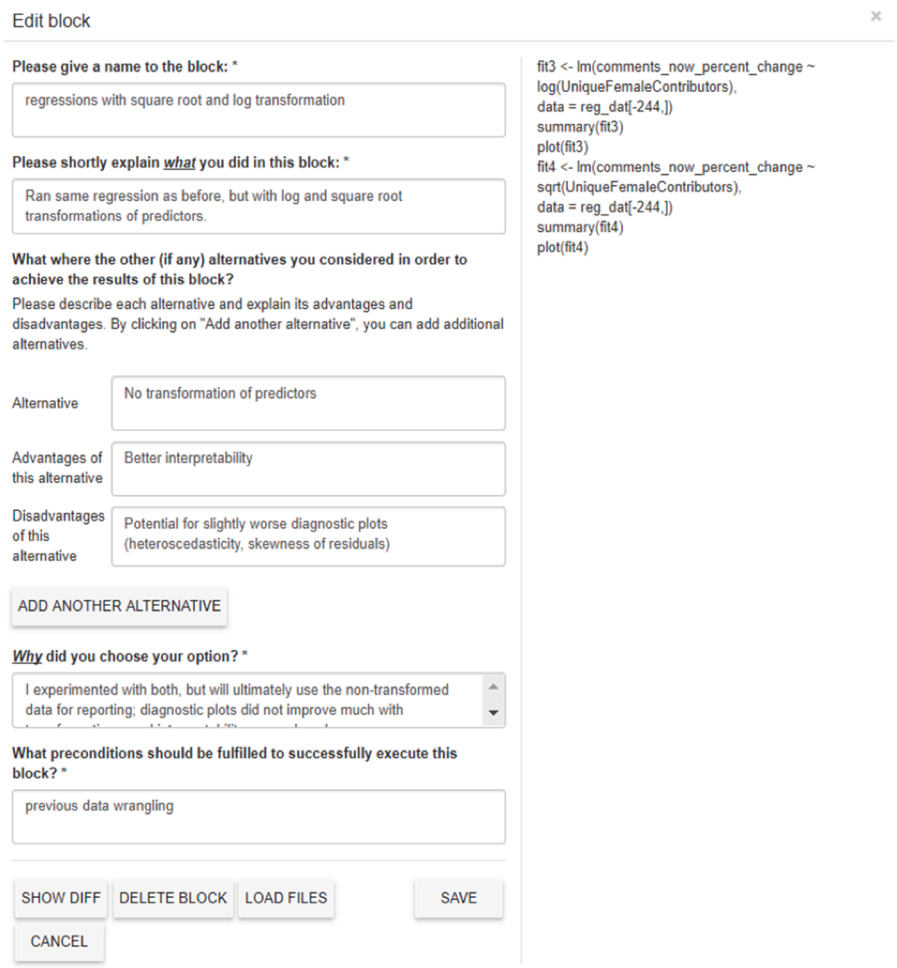


Fig. 1. Example block of logs with the explanations for the code.

Conversations consisted of two or more comments that were posted sequentially by at least one contributor. A contributor was one person who posted at least one comment to one or more conversations. Contributors often contributed several comments to the same conversation.

3.2. Recruitment and initial survey of analysts

Data analysts were recruited via open calls on social media platforms including Twitter, Facebook, forums of psychology interest groups, and R (R Core Team, 2018) mailing lists (see Supplement 4 for the project advertisements). In total, 49 scholars submitted analyses for this crowdsourcing initiative, of which 23 scholars completed 37 sufficiently detailed analysis reports (one report per hypothesis) and provided reproducible code suitable for inclusion. Notably, difficulties in reproducing analyses from the reported statistics (Bergh, Sharp, Aguinis, & Li, 2017), as well as the original data and code are common (Chang & Li, in press; Hardwicke et al., 2018; McCullough, McGeary, & Harrison, 2006; Stockemer, Koehler, & Lentz, 2018; Stodden, Seiler, & Ma, 2018), even under the most favorable of circumstances as with pre-registered reports (Obels, Lakens, Coles, Gottfried, & Green, in press).

Eight of the remaining analyses, from six analysts, were flagged by sub-teams of research assistants and independent statisticians as containing errors. See below and Supplement 7 and 8 for further details on the error and reproducibility checks, and the results of the excluded analyses. The overall rate of problems identified is not surprising since scientific errors are quite common (Bakker & Wicherts, 2011; Bergh et al., 2017; Rohrer et al., in press). The exclusions for errors left a total of 29 analyses, $N = 14$ for Hypothesis 1 and $N = 15$ for Hypothesis 2, which were conducted by 19 analysts, as the focus of this primary project report. The quantitative analyses below focus on these 29 results from 19 analysts.

Prior to receiving the dataset, analysts completed a pre-survey of their disciplinary background and expertise, and a set of demographic measures (see Supplement 5 for the complete pre-survey items and https://osf.io/y9fq4/ for the data). At the time of the project, participating analysts were on average 31.2 years of age ($SD = 7.2$), and included 15 men and 4 women. Seven resided in the United States, five in European countries, and the rest in Australia, Brazil, New Zealand, Pakistan, Russia, Singapore, and South Korea. Three were professors, one was a post-doctoral researcher, six were doctoral students, four held another academic position (e.g., data analyst), and five were not affiliated with an academic institution. The participating analysts self-reported an average of 6.5 years of experience in data analysis ($SD = 5.5$). A substantial minority indicated that they performed data analysis on a daily basis (7 analysts, 37%), while the rest performed data analysis a few times a week (3 analysts, 16%), once a week (4 analysts, 21%), once every two weeks (1 analyst, 5%), or less (4 analysts, 21%).

3.3. Analyses using the DataExplained platform

We designed an online platform called DataExplained that supports transparent data analysis reporting in real time. The platform records all executed source code and prompts analysts to comment on their code and analytical thinking steps. DataExplained is based on RStudio Server

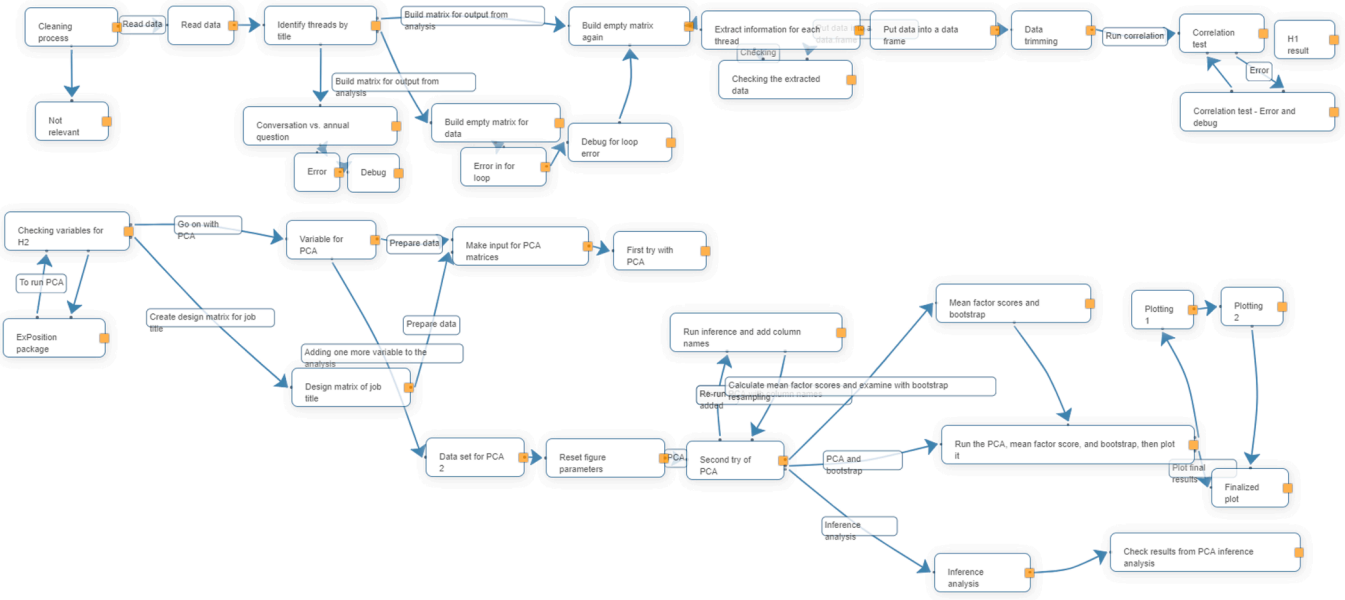


Fig. 2. Snippet of workflow modeled by a participating analyst.

(https://www.rstudio.com/products/rstudio-server/), a data analysis platform that allows users to conduct analyses remotely via a web browser based on the familiar RStudio interface. In addition to the online RStudio environment, we implemented features that enabled us to track all executed commands along with the analysts' detailed explanations for every step of the executed analysis.

The procedure was as follows. First, the participants were provided access to the platform, where they executed their data analysis using the RStudio user web-interface. During their analysis, every executed command (i.e., log) was recorded. Recording all executed commands (including commands executed but not necessarily found in the final code) is useful, as such logs might reveal information that affected the analysts' decisions but are not reflected in the final script. Whenever the participants believed that a series of logs could be described as a self-explanatory block, or when a certain number of logs was produced, they were asked to describe their rationales and thoughts about the underlying code. The dataset was available in the environment of DataExplained only. Use of this platform essentially involves conducting analyses in R with added transparency features.

We included a number of elements to capture the workflow of analysts. In particular, once the analysts reached a certain number of executed commands, we prompted them to explain the goals and reasoning underlying the relevant code, as well as alternative approaches they rejected. As shown in Figure 1, this consisted of a few key questions: 1) *Please shortly explain what you did in this block?*, 2) *What preconditions should be fulfilled to successfully execute this block?*, 3) *What were the other (if any) alternatives you considered in order to achieve the results of this block? (explain the alternative, explain the advantages, explain the disadvantage)*, and 4) *Why did you choose your option?* This allowed us to observe the reasons underlying an analytic decision, the justification for it, the considered alternatives, the trade-offs evaluated, and the deliberation that led to the final implementation.

To provide a useful unit of analysis, we asked the analysts participating in our study to split workflows (i.e., the whole sequence of all commands used in the analysis) into semantic blocks (essentially, subsequences of commands). This way, each block was annotated with descriptive properties which reflect the rationales and reasoning of the analyst's actions within a block. Analysts were able to navigate through their analysis history, by restoring the state of the RStudio workspace at any given point a block was created. These features helped the analysts to recall the considerations during their analysis, even if the corresponding portion of code was no longer in the final script.

Finally, DataExplained provided analysts with an overview of all blocks that they created and asked them to graphically model the workflow representing the evolution of the analysis. Initially, each analyst was presented with a straight chain of blocks, ordered by their execution. The analysts were then asked to restructure the workflow such that it better reflected their actual process. For example, iterative cycles of trying out different approaches for a sub-problem could be modeled as loops in the workflow. Figure 2 shows an example workflow visualization from an analyst in the present crowdsourced project. The orange boxes displayed in Figure 2 allowed analysts to connect the various steps of their analysis. Clicking on an orange box produced an arrow, which could then be connected to any other of the analysts' steps. For example, an analyst who wanted to indicate that "Step A" led her to "Step B" would first click on the orange box of "Step A" and then drag the resulting arrow to "Step B." A video demonstration of this process is available at https://goo.gl/rnpgae, see in particular minute 04:30 for how steps are linked.

3.4. Post-survey

After completing their analyses via the DataExplained platform, analysts responded to a second survey in which they were asked to report their empirical results and the analytic methods they used, such as transformations, exclusions, statistical techniques, covariates, and operationalizations (see Supplement 6 for the complete post-survey and https://osf.io/u8rmw/ for the data).

3.5. Independent assessment of analysis quality

Finally, two teams of research assistants and statisticians carefully reviewed each analyst's approach for errors and ensured they could independently reproduce the results (see Supplements 7 and 8 and htt

Table 1.1
Overview of analytic approaches and results across independent scientists for Hypothesis 1, "A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion"

Analyst*	Statistical approach	Sample size	Unit of analysis	Covariates	Operationalization of female participation in academic discussions	Operationalization of number of women in discussion	Effect size
1	logistic regression	5443	Comments	None	odds of next contributor to conversation being a woman	cumulative sum of previous female comments in a conversation	1.06 odds ratio
2	linear regression	65	combination of conversations and proxy for number of contributors	None	proxy for number of comments by each female contributor in a conversation	number of female contributors ordered by time of commenting (first, second, third female contributor, etc)	−1.32 regression coefficient
3	generalized linear mixed effects regression (Poisson)[1]	645	Comments	number of comments in a conversation	number of comments by author in a conversation (females only)	percentage of unique female contributors in a conversation	0.33 regression coefficient
4	Pearson correlation	7975	Comments	None	number of comments made by all female contributors in a conversation	number of unique female contributors in a conversation	0.87 correlation coefficient
5	Pearson correlation	270	Comments	None	number of comments made by all female contributors in a conversation	percentage of comments made by females in a conversation	0.56 correlation coefficient
6	linear regression	462	combination of conversations and contributors	None	difference between female comments in current conversation and previous conversation	number of unique female contributors in a conversation	−0.59 regression coefficient
7	logistic regression	4502	Comments	academic discipline	whether the current contributor is a woman	cumulative sum of female comments that precede a specific comment	0.15 regression coefficient
9	linear regression	634	Comments	None	number of words in a female comment	cumulative proportion of female comments in each conversation	23.47 regression coefficient
11	generalized linear mixed effects regression (Poisson)[2]	463	combination of conversations and contributors	None	number of comments by author in a conversation (females only)	number of unique female contributors in a conversation	−0.02 regression coefficient
12	generalized linear regression (Poisson)	96	Conversations	1) debate size 2) conversation written / transcribed	number of comments made by all female contributors in a conversation	percentage of unique female contributors in a conversation	27.3 incidence rate ratio
13	linear regression	504	Conversations	total number of unique contributors in a conversation	percentage of comments made by women in a conversation	number of unique female contributors in a conversation	0.26 regression coefficient
14	linear regression	36	Conversations	None	percentage of comments made by women in a conversation	number of unique female contributors in a conversation	−0.001 regression coefficient
17	Kendall correlation	96	Conversations	None	proxy for average number of comments made by each woman in a conversation	percentage of unique female contributors in a conversation	0.37 correlation coefficient
19	linear regression	193	Comments	1) number of prior comments, 2) contributor has PhD/ not, 3) total citations	number of comments by author in a conversation (females only)	number of unique female contributors in a conversation	−0.32 regression coefficient

Notes. This table includes analyses not flagged as having clear errors by independent reviewers.
This table includes the original effect sizes reported by the analysts, which are not directly comparable to one another.
* In the online article, the column includes hyperlinks for each analyst's error checks and raw code
[1] Random intercept for conversation ID; random intercept and slope for contributor ID
[2] Random intercept for conversation ID

ps://osf.io/n5q3c/). These error-checks involved a two-step process. First, three research assistants from The European School of Management and Technology (ESMT) conducted an initial review and error check. These three RAs were graduate students in computational neuroscience, public policy, and economics and were selected for their strong data analysis backgrounds. They had advanced knowledge of statistics and econometrics and were skilled in R, Python, Matlab, and Stata. Two of the ESMT research assistants coded each analysis for potential errors, and if they found any discussed this with each other to clarify whether they agreed on an analytical choice being an error or not. If need be, they also consulted a third ESMT research assistant and/or the first author. The RAs created an error check document for each analysis which contained the entire code, a summary of the code, key information about each analysis, and an indication whether they suspected any serious errors. Second, a team of statistical experts based at the Tilburg University Department of Methodology (a graduate student,

Table 1.2
Overview of analytic approaches and results across independent scientists for Hypothesis 2, "Higher status participants are more verbose than lower status participants"

Analyst*	Statistical approach	Sample size	Unit of analysis	Covariates	Operationalization of verbosity	Operationalization of status	Effect size
1	linear regression	4262	Comments	1) contributor gender 2) contributor in academia or not	number of characters in a comment	academic job rank (postdoc, professor, etc…)	−0.16 regression coefficient
3	linear mixed effects regression[1]	1497	Comments	1) academic job rank 2) university ranking	number of words in a comment	total number of citations	0.04 regression coefficient
5	linear regression	306	Comments	None	number of conversations in which a contributor has participated in a specific year	job title	3.97 regression coefficient
6	linear regression	297	Contributors	None	average number of words in a conversation	academic job rank	−64.38 regression coefficient
7	linear regression	1537	Comments	1) academic job rank 2) discipline	number of characters in a comment	total number of citations	−0.22 regression coefficient
9	linear regression	721	Contributors	None	average number of words in all comments	combination of: 1) whether a contributor has a PhD or not and 2) rank of their academic workplace	69.70 regression coefficient
10	linear mixed effects regression[2]	7718	Comments	1) contributor gender 2) contributor role (author or commentator) 3) type of exchange (annual questions or conversations)	number of words in a comment	combination of: whether a contributor has a PhD or not, whether a contributor is in academia or not, the rank of their PhD institution and academic workplace, total number of citations, academic job rank, and the number of conversations in which a contributor has participated	0.12 regression coefficient
11	linear mixed effects regression[3]	857	Comments	1) contributor gender 2) number of citations 3) academic job rank 4) number of years since received PhD	number of words in sentences	h-index	0.09 regression coefficient
12	linear regression	1007	combination of contributors and status-related variables	1) contributor gender 2) discipline	average number of words in all comments	academic job rank	54.39 regression coefficient
14	linear mixed effects regression[2]	518	Comments	1) total number of citations 2) university ranking	number of characters in a comment	rank of contributor's academic workplace where higher values indicate lower rank	0.06 regression coefficient
17	Kendall correlation	4263	Comments	None	number of words in a comment	academic job rank	−0.05 correlation coefficient
18	linear mixed effects regression[2]	573	combination of contributors and conversations	collection of variables that include gender, whether the person is the first to contribute, conversation year, conversation type, and interaction terms between them	proxy for the number of characters, and the number of times a person contributes to the conversation	proxy for the combination of: 1) academic job rank and 2) the year when PhD was obtained	0.13 regression coefficient
21	factorial ANOVA, Eta-squared value	355	Contributors	None	average number of words in all comments	academic job rank	0.02 eta squared
22	Spearman correlation	728	Contributors	None	number of comments in a year	academic job rank	−0.04 correlation coefficient
23	linear regression	386	combination of contributors and academic job rank	contributor gender	average number of characters in all comments	academic job rank	−239.01 regression coefficient

Notes. This table includes analyses not flagged as having clear errors by independent reviewers.
This table includes the original effect sizes reported by the analysts, which are not directly comparable to one another.
* In the online article, the column includes hyperlinks for each analyst's error checks and raw code
[1] Random intercept for contributor ID; random intercept and slope for conversation ID
[2] Random intercepts for conversation ID and contributor ID
[3] Random intercept for whether the conversation was written / transcribed

postdoctoral researcher, and professor) reviewed these error checks and individual analyses, again examining whether the code by each analyst contained any serious errors. The error check documents are publicly posted at https://osf.io/n5q3c/. In the end the ESMT and Tilburg subteams converged on a subset of analyses that were deemed as containing errors. As noted earlier, only error-free and fully reproducible analyses ($N = 14$ for Hypothesis 1 and $N = 15$ for Hypothesis 2) are included in this primary report of the quantitative results. The results with excluded analyses are provided in Supplement 7.

4. Results

4.1. Variability in analytic approaches and conclusions

We set out to identify the extent of heterogeneity in researchers' choices of analytic methods, and the impact of this heterogeneity on the conclusions drawn about research questions regarding gender and professional status in group meetings. We found that the participating analysts employed a wide array of statistical techniques, covariates, and operationalizations of key theoretical variables such as professional status and verbosity (see https://osf.io/n5q3c/ for the code for each individual analyst). As summarized in Tables 1.1–1.3, different analysts operationalized variables in various ways: for example, Analysts 3, 10, and 17 operationalized verbosity as the number of words contributed in a comment, Analyst 5 operationalized verbosity as the number of conversations participated in, and Analysts 1, 7, and 14 operationalized verbosity as the number of characters in comments, among other approaches. Status was assessed using academic job rank, citation count, h-index, and university rank, as well as via a combination of indicators. Additionally, the unit of analysis varied. For example, Analyst 9 in H1 focused their analyses on the level of comments by counting the number of words in a comment made by a female contributor, whereas Analyst 12 focused their analyses on the level of conversations by counting the number of comments made by all female contributors in a conversation. Sample size varied greatly even for analyses on the same unit of analysis. Strikingly, no two individual analysts employed precisely the same specification for either Hypothesis 1 or 2 (see Botvinik-Nezer et al., 2020, and Carp, 2012a; 2012b, for similar findings in neuroimaging studies and Bastiaansen et al., 2020, for a conceptual replication with event sampling data from a clinical patient).

The crowd of independent researchers further obtained widely varying empirical results regarding Hypothesis 1 and 2, using widely varying statistical techniques, and reported statistically significant results in both directions for each hypothesis. Table 2 summarizes the number of analysts who obtained statistically significant support for the hypothesis, directional but non-significant support, directional results contrary to the hypothesis, and statistically significant results contrary to the initial prediction. As seen in the table, while 64.3% of analysts reported statistically significant support for Hypothesis 1, 21.4% of analysts reported a statistically significant effect in the opposite direction (i.e., finding that a woman is *less* likely to contribute to the conversation when there are other women in the meeting). At the same time, while 28.6% of analysts reported significant support for Hypothesis 2, 21.4% reported a significant effect in the contrary direction (i.e., finding that high status participants are *less* verbose that lower status participants).

Although we do not defend the use of *p*-value cutoffs for deciding what is true and what is not, a reliance on such thresholds by both authors and gatekeepers (e.g., editors and reviewers) is extremely common in the fields of management and psychology (Aguinis et al., 2010). Thus, Table 2 does give us a sense of what might have been published had a single analyst conducted the research alone. In other words, had a crowdsourced approach *not* been employed, there would have been a roughly 1 in 4 chance of a research report of statistically significant support for Hypothesis 2, about a 1 in 4 chance of a report of the opposite pattern, and a 2 in 4 chance of null results. Further, in all of these scenarios, the role of subjective researcher decisions in the published outcome would have remained unknown rather than made transparent.

4.2. Dispersion in standardized scores

Given the diversity in analytical choices and approaches, it is not straightforward to compare or aggregate all the results. Tables 1.1 and 1.2 include the effect size estimates reported by the individual analysts, which are not directly comparable to one another. We encountered two challenges when attempting to compute standardized effect sizes on the same scale for all independent analyses of the same hypothesis. First, most analyses were non-standard, so we often lacked a well-known and commonly used effect size measure. Second, even after applying or developing specialized effect size measures, there is no means by which to convert all these different effect sizes to the same effect size metric. We bypassed these problems by computing the *z*-score for each statistical result's *p*-value, which is also done before analyzing data in Stouffer's method in *meta*-analysis and *z*-curve (Brunner & Schimmack, 2018). This method transforms individual *p*-values of test statistics to *z*-scores, assuming that the sampling distribution of the test statistic is approximately normally distributed, resulting in random variables with a variance of 1.

It is crucial to realize that the analysts' *z*-statistics are a function of the effect size, the number of independent observations in the analysis, as well as the selected statistical technique and their statistical properties (e.g., statistical power, in case of a true nonzero effect). As the three aforementioned factors are all affected by the analysts' selected analysis, and all analysts use the same dataset, differences in *z*-scores still reflect differences in the consequences of analysts' choices.

Regarding the normality assumption of the *z*-scores, note that most parameters in models correspond to linear combinations of the data. For instance, a mean or probability (sum of values divided by N), variance (sum of squared deviations divided by N-1), a regression coefficient (sum of (X-Xmean)*(Y-Ymean) divided by a constant equal to (X-Xmean)^2). If the sum is over independent observations, then it follows from the central limit theorem that all these sums are increasingly better approximated by the normal distribution for larger N. More generally, many test statistics are well approximated by a normal distribution for larger N. Except for the z-statistics, think of the t-statistic (same shape but a bit larger variance), the Chi2-statistic (similar shape but skewed to the right), and for the F-statistic but only when df1 = 1 (this is the t) or when df1 has a 'large' value. Tables 1.1 and 1.2 contain detailed information about the number of observations used in the analyses. For example, Analyst 1 for H1 drew on a sample of 5,443 observations. The sample sizes for all other analyses are reported in these tables. As most statistics are well approximated by a normal distribution for the number of observations considered by the analysts, we believe that the normal approximation works rather well in this application.

The *z*-scores of individual results were obtained using different methods. In some cases the *z*-scores could be directly retrieved from the output of the analyst, but in the majority of the cases *z*-scores were

Table 1.3
Breakdown of choice points and approaches for each hypothesis tested.

Choice point	Hypothesis 1	Hypothesis 2
Independent variable	64% of analysts operationalized "number of women in discussion" as the number/percentage of unique female contributors in a conversation, 21% as the cumulative sum/proportion of female comments that preceded a specific comment, 7% as the percentage of comments made by women in a discussion, and 7% as the number of female contributors ordered by time of commenting.	47% of analysts operationalized "status" as contributor's academic job rank, 13% as total number of citations, 7% as H-index, 7% as rank of the academic workplace, 7% as job title, and 20% as a combination of different status-related variables.
Dependent variable	57% of analysts operationalized "female participation in academic discussions" as number of comments made by female contributors in a conversation, 14% used percentage of comments made by women, 7% as the number of words in comments from women, 7% as the odds of the next contributor to a conversation being a woman, 7% as whether the current contributor is a woman or not, and 7% as the difference between the number of female comments in previous and current conversations.	47% used number of words in comments / conversations to operationalize "verbosity", 27% used number of characters in contributor's comments, 7% used number of comments a contributor made in a year, 7% used number of words in sentences, 7% used number of conversations in which a contributor has participated in a specific year, and 7% used a combination of number of characters in comments and number of times a person contributes to a conversation.
Covariates	64% did not use any covariates, 7% used number of comments in a conversation, 7% academic discipline, 7% total number of unique contributors in a conversation, 7% debate size and whether the conversation was written or transcribed, and 7% used a combination of variables that included number of prior comments for a contributor, whether the contributor has PhD or not, and contributor's total number of citations.	40% did not use any covariates, 7% used contributor's gender, 7% used contributor's gender and whether the contributor is in academia or not, 7% used contributor's academic job rank and their university ranking, 7% used contributor's job rank and their discipline, 7% used contributor's gender and discipline, 7% used contributor's total number of citations and their university's ranking, and 20% used a combination of contributor-related variables such as gender, number of years since PhD obtained, and role in the conversation.
Unit of analysis	50% of analysts chose comments as their unit of analysis, 29% chose conversations, 14% chose a combination of conversations and contributors, and 7% created a custom unit of analysis as a combination of conversations and a proxy for the number of female contributors.	53% of analysts chose comments as their unit of analysis, 27% chose contributors, 7% chose a combination of conversations and contributors, 7% created a custom unit of analysis as a combination of contributors and status-related variables, and 7% as a combination of contributors and academic job rank.
Statistical approach	43% used linear regression to analyze the data, 14% opted for logistic regression, 14% chose generalized linear mixed effects regression, 14% Pearson correlation, 7% Kendall correlation, and 7% generalized linear regression.	47% decided on linear regression to analyze the data, 33% opted for linear mixed effects regression, 7% Spearman correlation, 7% Kendall correlation, and 7% factorial ANOVA.

computed using the *p*-value of the test statistic (using the quantile normal distribution in R). In one case where a *p*-value was not presented by the analyst we ran our own code in R to retrieve it (i.e., cor.test(data2 $TendencyToParticipate, data2$UniqueFemaleContributors, method-="kendall"). Sometimes a large *t*-value was provided in combination with its *df* and a *p*-value < 0.001. In those cases, the exact *p*-value was first calculated using R, and then transformed to a *z*-score (e.g., $t(100) = 10$ is transformed to $z = 8.306$ by qnorm(pt(10,100, lower.tail = FALSE), lower.tail = FALSE)). As *t*-values could be very large or *p*-values very small, we sometimes had to use the log.p argument to obtain *z*-values (e.g., $t(7000) = 100$ results in −3,110.64 using pt(100,7000, lower.tail = FALSE,log.p = TRUE), which yields $z = 78.81$ using qnorm (-3,110.64,lower.tail = FALSE,log.p = TRUE). Finally, it was possible to compute a *z*-score from the 95% confidence interval of a result (e.g., an estimate = *x* and lower bound = *y* yield *z* <- *x*/((*x*-*y*)/qnorm(0.975))). See r file "*specification_curve_2.R*" and Excel file "*ES Transformations 2.1 anonymized IDs_140120.csv*" for details on how the specification curve analyses (Simonsohn et al., 2020) were conducted (https://osf.io/fgrjq/).

Figures 3 and 4 display the results reported by the different analysts after converting them to standardized scores, and further provides some details on the analytic approaches employed (following on Simonsohn et al., 2020). The *z*-scores corresponding to the estimate for Hypothesis 1 ranged from −7.230 to 106.267, with a median of 7.027, and mean of 12.329 (standard error = 0.267) that was significantly different from zero ($z = 46.131$, two-tailed $p < .001$). The *z*-score corresponding to the estimate for Hypothesis 2 ranged from −4.394 to 7.450, with a median of 0.700, and mean of 0.685 (standard error = 0.258), which was also significantly different from zero ($z = 2.653$, two-tailed $p = .008$). That the means differ from zero is less informative as for both hypotheses some analysts found the opposite result (i.e., a negative effect). Evidence of an effect is stronger for Hypothesis 1 than for Hypothesis 2, which is signified by the larger Spearman rank order correlation between absolute *z*-score and sample size for Hypothesis 1 ($r_s = 0.689$, one-tailed $p = .003$) than for Hypothesis 2 ($r_s = 0.364$, one-tailed $p = .091$). The standardized scores were heterogeneous for both Hypothesis 1 ($\chi^2(13) = 10{,}171.57$, $p < .001$) and Hypothesis 2 ($\chi^2(14) = 165.73$, $p < .001$), confirming the greatly diverging analyses and their outcomes.

4.3. Qualitative coding of quantitative analytic decisions

That cognitive processes play a key role in data analysis has been

Table 2
Direction and significance levels for results from the independent analysts for Hypothesis 1 and Hypothesis 2.

Hypothesis	Significant in predicted (+) direction	Not significant in predicted (+) direction	Not significant in opposite (-) direction	Significant in opposite (-) direction
H1: A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion	64.3% ($n = 9$)	0% ($n = 0$)	14.2% ($n = 2$)	21.4% ($n = 3$)
H2: Higher status participants are more verbose than lower status participants	28.6% ($n = 4$)	21.4% ($n = 3$)	28.6% ($n = 4$)	21.4% ($n = 3$)

Note. For Hypothesis 2, analyst 21 found a non-directional, nonsignificant effect (eta squared). Only those analyses are included in this table for which both direction and significance levels were known (i.e., for H1: analysts 1, 2, 3, 4, 5, 6, 7, 9, 11, 12, 13, 14, 17, 19 and for H2: analysts 1, 3, 5, 6, 7, 9, 10, 11, 12, 14, 17, 18, 22, 23).

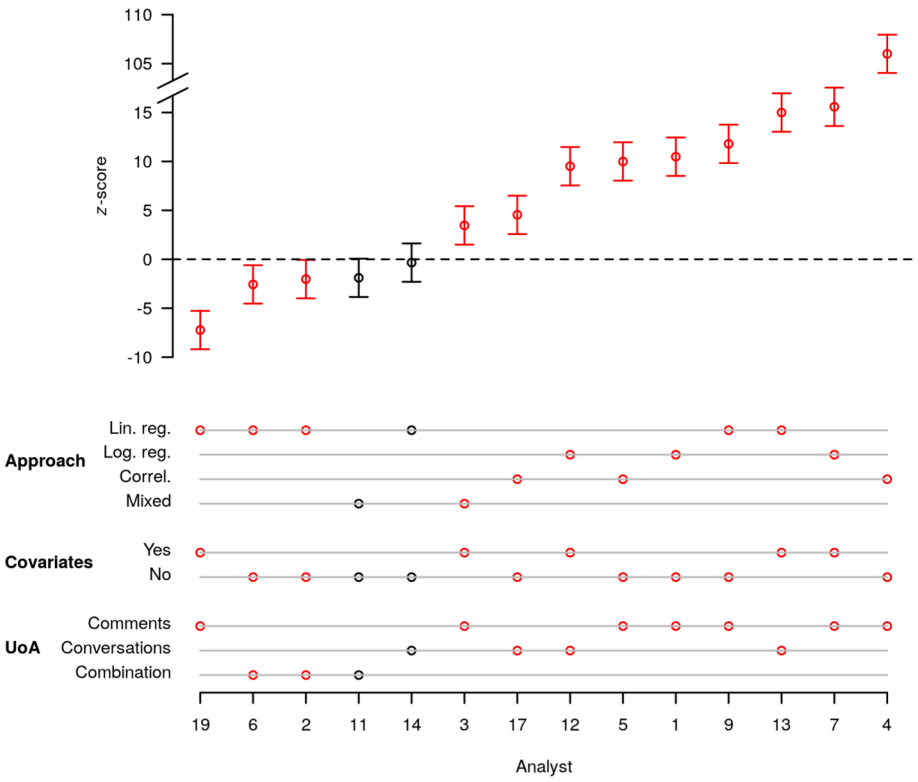


Fig. 3. Dispersion of z-scores corresponding to estimates of independent analysts using the same dataset to test Hypothesis 1 ("*A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion*"), together with some details on each specification. Note that there is a break in the y-axis of the figure to incorporate the extreme z-score of Analyst 4.

acknowledged for many years by statisticians (Tukey & Wilk, 1966). The process of building and interpreting the relevant mental models or schemas is known as sensemaking. Weick, Sutcliffe, and Obstfeld (2005) define sensemaking as "the ongoing retrospective development of plausible images that rationalize what people are doing" (p. 409). Through DataExplained, we are able to observe the roadmap of different analytical alternatives and justifications for decisions in much greater detail than ever before. To better understand the sensemaking process underlying these analytic decisions, we relied on a qualitative research approach. A project sub-team of qualitative researchers analyzed the descriptive text explaining in detail every step undertaken by individual analysts throughout their data analyses as well as the source-code corresponding to each step.

By asking analysts to explain their decisions and considered alternatives to the executed code, we obtained a rich dataset capturing their various workflows. This is especially useful due to the exploratory element of data analysis, where researchers often experiment with data prior to deciding on how to proceed. Indeed, graphic representations of the analysts' R-codes show that the analyses were often iterative, seemingly lacked a clear direction at times and instead included several explorative loops which help analysts make sense of the data over time. The relatively unstructured nature of the R-codes provided did not facilitate quantitative numeric or quantitative text analyses. Instead we decided to use the General Inductive Approach (Thomas, 2006) because this allowed us to analyze the R-code from the bottom up, subjecting each line of code to an iterative, qualitative analysis. This qualitative approach helped us understand how analysts made sense of the data and the factors guiding their decision-making processes. The goal of this approach is to translate qualitative raw data describing a process or experience into a consistent behavioral model reflecting a latent structure driving the process described in the text data.

Inductive coding is central to the General Inductive Approach. Our process began with multiple coders carefully reading the relevant materials and considering possible meanings reflected in the text. Below, by "researchers" we refer to the independent analysts participating in the crowd project, and by "coders" we mean the separate sub-team organized to carry out the meta-scientific qualitative analyses of the crowd analysts' quantitative decisions. The team of qualitative coders identified text snippets that contained meaningful information and created *codes* (i.e., labels or tags) best describing the main insight of the snippet. After the coders refined a set of codes, they developed an initial description of the meaning of each code along with a *memo* – a short description explaining the code and elaborating on when it should be applied. Eventually, the codes from different coders were merged and discussed as a group. All codes as well as their memos were aggregated together into a code book, provided in Supplement 9 (see also Feldman, 2018, and Staub, 2017). The coders then iteratively kept refining and re-evaluating the codebook until the process reached a well-established and shared understanding of all the codes (see Figure 5).

A detailed report of this bottom-up qualitative analyses of the annotated code from DataExplained is provided in Feldman (2018), Staub (2017), and in Supplement 9. Our analytical approach was bottom-up in that we qualitatively analyzed individual blocks of code. Specifically, we closely read the analysts' blocks of code, as well as their responses to open questions about their analytical choices such as: "Why did you choose your option?". Following the General Inductive Approach (Thomas, 2006) we identified meaningful units in these responses and assigned different labels to these meaningful units. For example, if an

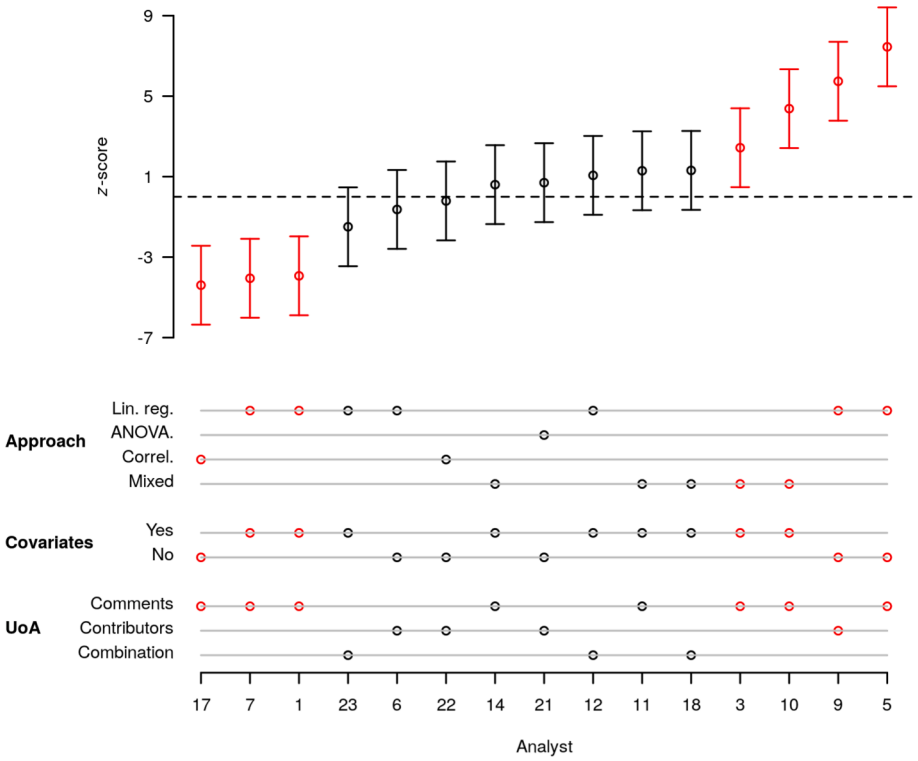


Fig. 4. Dispersion of *z*-scores corresponding to estimates of independent analysts using the same dataset to test Hypothesis 2 (*"Higher status participants are more verbose than lower status participants"*), together with some details on each specification.

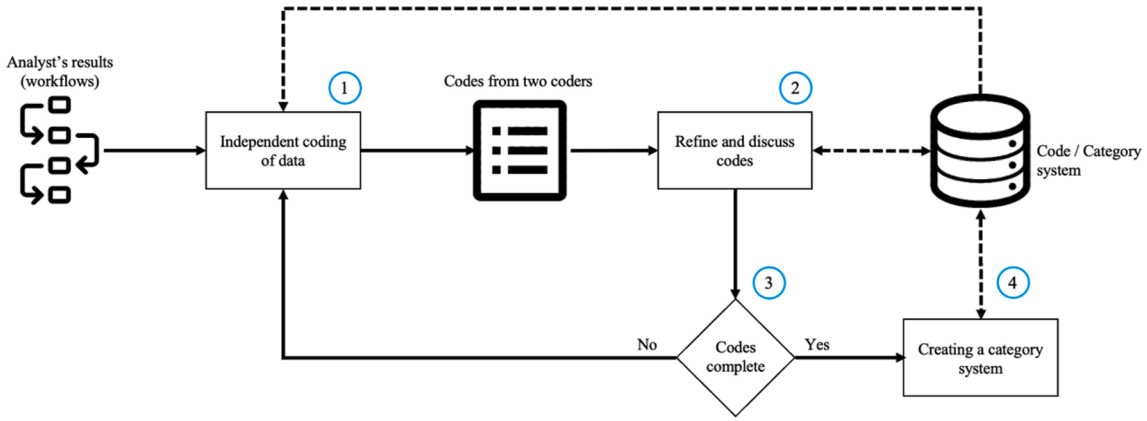


Fig. 5. The workflow of our qualitative analysis of the quantitative analytic decisions.

analyst responded "I experimented with both, but will ultimately use the non-transformed data for reporting; diagnostic plots did not improve much with transformations, and interpretability was reduced", we assigned the label "exploratory" to this response. Over time, and over coding many of these responses, meaningful categories, or "key factors" emerged, which seemingly influenced analytical choices analysts made.

In order to ensure the reliability of the emerging codes and categories, we applied both qualitative and quantitative measures of reliability (Campbell, Quincy, Osserman, & Pedersen, 2013; Kurasaki, 2000; Hruschka et al., 2004). Two coders followed multiple coding cycles (see Figure 5) in order to build a sustainable coding scheme. The proportional agreement of the two coders after the last iteration was 72%, with a Cohen's Kappa of 0.70. The resulting codebook was then presented to two new coders. After further iterations performed by all four coders, the percentage agreement reached 52.6%. The team of coders identified patterns in researchers' reasoning (about data constraints, preprocessing steps, the hypothesis, alternative methods, etc.) using the final set of 31 codes grouped into 10 categories and 4 meta-categories.

These codings led to a proposed model of the data analyst's reasoning process and workflow (Figure 6). The model seeks to capture the iterative interplay between understandings of the dataset and hypotheses to be tested, the analyst's knowledge and beliefs, the actions and methods actually performed during the analysis, and insights gained. As researchers conduct data analyses, they obtain intermediate results. These results are almost always interpretative in their nature and often stem from personal understanding and beliefs, which often vary across individuals. Data analysis is an iterative process, and intermediate output plays a key role in deciding which path to further follow. The data by itself can influence an analyst's beliefs, which as a consequence

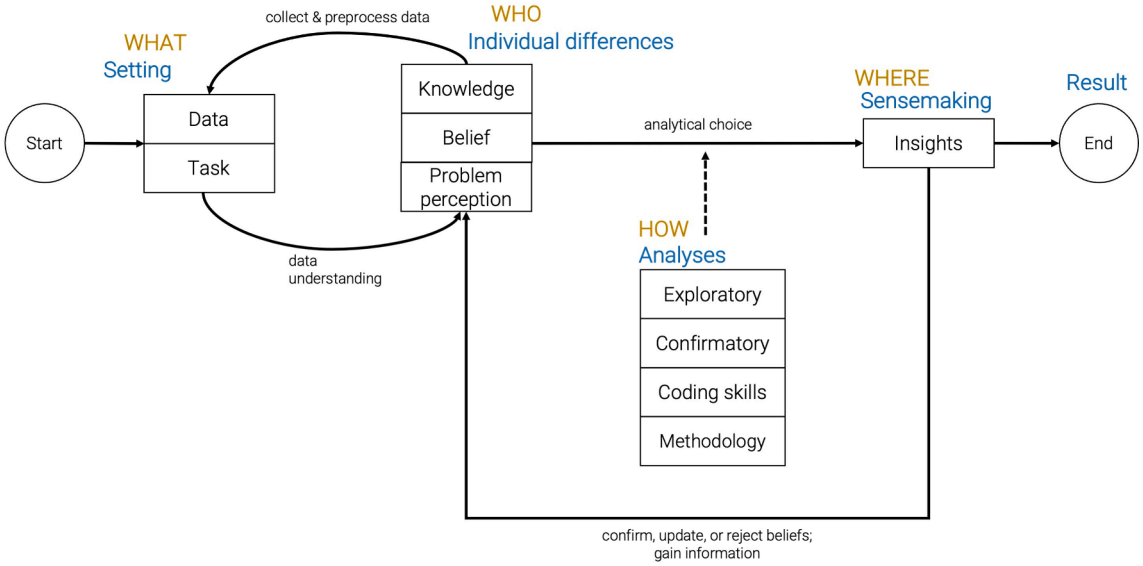


Fig. 6. Model of an analyst's reasoning process.

may lead to different analytical choices. Thereby, a data analysis not only incorporates statistical or computational steps, but also cognitive processes (Grolemund & Wickham, 2014; Paglieri, 2004). The four meta-categories derived from our qualitative coding form the core of a model of the cognitive processes involved in data analysis.

What (setting). This meta-category covers the elements of the process which are given and objective in nature. The dataset structure and characteristics and (for this crowdsourced project) the specific hypothesis they are tasked with testing are the same for different data analysts. The sub-categories under this *meta*-category are *Data* and *Task*. Note that these elements might still be interpreted in various ways (e.g., due to new insights or personal beliefs), but cannot be changed. Having data and task (e.g., hypothesis to test) at hand, the analyst then proceeds to understand the data. This process of understanding is where the first source of subjectivity can be observed due to differences between analysts.

Who (personal). The second meta-category relates to personal attributes of the data analyst. This includes the sub-categories *Knowledge*, *Beliefs,* and *Problem perception* which reflect the contribution of personal attitudes and biases in problem-solving in general as well as in data analysis. Even the way data is preprocessed (cleaned, subsampled, aggregated etc.) can be a consequence of person factors, leading to variability.

How (analysis). The "how" meta-category captures actions or methods which are performed during data analysis. These can either be exploratory or confirmatory in nature. We refer to exploratory data analysis (EDA) as the process of data exploration, as well as attempts to understand the logic of the problem and summarize its main characteristics. Confirmatory data analysis (CDA) refers to the analytic choices to confirm the emerged models (i.e., systematically assess the strength of evidence). Note that this is a different definition of a confirmatory analysis than seen in scholarship on pre-registration of analyses, in which strictly confirmatory analyses are planned out and "frozen" online prior to having the dataset (Wagenmakers et al., 2012).

Where (sensemaking). Data analysis can be an iterative process where each iteration leads to new insights gained. The "Where" or sensemaking meta-category is the point at which the analyst processes the results of the previous iteration and makes a decision on how to proceed. The analyst decides whether to confirm, update, or reject her or his current understanding of the problem due to insights gained from the previous iteration. These underlying assumptions and beliefs help analysts determine where to allocate more attention and how to interpret the data (Klein, Moon, & Hoffman, 2006). Information that does not match pre-existing schemas may be overlooked or explained away, but can also be updated if the signal coming from the data is especially strong.

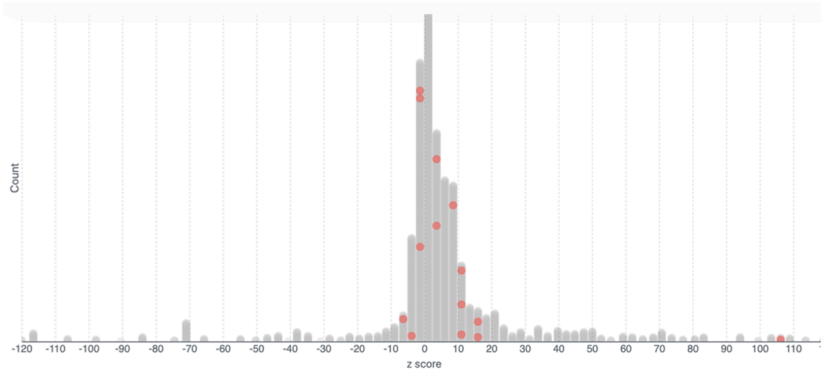
In the model, the initial specifications of the data analysis task as well as the data at hand (i.e., "WHAT"), interact with the prior beliefs and understandings of the person performing the analysis (i.e., "WHO"). The analyst's beliefs, accumulated knowledge, and past experiences impact problem perception and the way the data is interpreted. At the same time, the data is often reshaped and prefiltered in a way that is in harmony with the prior beliefs of the analyst. Further, analysis of the data can be seen as a spiral-like process where each iteration leads to new insights. As a result, an analyst makes decisions on how to proceed with her data analysis and advances further in a certain direction (i.e., "WHERE"). During this process, the analyst decides whether to confirm, update or reject her current understanding of the problem due to insights gained from the previous iteration. Since the way data analysis is carried out influences the final results (i.e., "HOW"), we describe variables such as methodology, codings, and exploratory and confirmatory data analysis as factors influencing the final empirical results of the research. In a series of iterative loops, analysts engage in this ongoing retrospective development to build and interpret mental models and schemas that make sense of the data they are confronted with. The model in Figure 6 was empirically derived and, to the best of our knowledge, is the first to provide a detailed, data grounded overview of the behavioral factors involved in the data analysis process.

In harmony with these qualitative findings regarding the subjective sense-making process underlying data analysis, the quantitative results demonstrate that researchers ultimately select a wide variety of operationalizations of variables and statistical approaches, leading to radical dispersion in empirical findings (Tables 1.1, 1.2, 1.3 and Table 2, and Figures 3 and 4). Of course, our quantitative and qualitative *meta*-scientific analyses of the project results are no doubt affected by subjective researcher decisions as well. In the spirit of crowdsourcing, we welcome alternative perspectives on the publicly posted data from this initiative.

4.4. Boba multiverse analysis

To complement the qualitative analyses based on DataExplained, we also examined underlying processes quantitatively, through a Boba multiverse analysis (Liu et al., 2020). This crossed all of the crowd of analysts' choices with one another, removing analytic choices that did not make sense in conjunction with one another (e.g., apply logistic

a) *z-scores for Hypothesis 1:*



b) *z-scores for Hypothesis 2:*

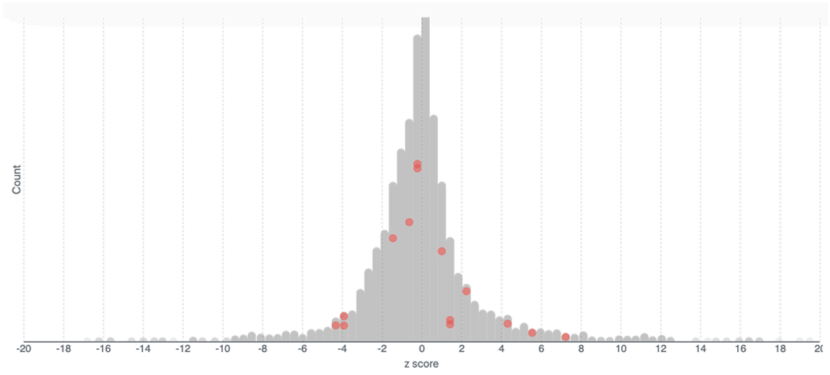


Fig. 7. In the Boba multiverse analysis, z-scores for Hypotheses 1 and 2. Outcomes from the crowd analysts are highlighted in red and represent only a subset of the multiverse of possible analyses.

regression analysis to a continuous dependent variable), or instances in which the independent and dependent variable would have been identical (e.g., percentage of comments made by females was used as independent variable by some analysts and as dependent variable by other analysts). We also excluded paths that produced run-time errors. As seen in Figure 7, top panel, the majority of z-scores are positive for H1, suggesting an overall positive effect. In contrast, H2 seems to be quite symmetrical around zero, suggesting no effect or a tiny effect.

The Boba multiverse approach allows us to parse some of the contributors to dispersion of estimates, identifying some of the key steps in this garden of forking paths (Figure 8). More specifically, we examined how different analytic choices were associated to the outcome of an analysis. We used two methods to do this, each focusing on a slightly different question. The first method utilizes adjusted R^2 to quantify the variance explained by any analytic choice or any combination of two analytic choices. To obtain the adjusted R^2, we fit a linear model where we used one choice or two choices and their interaction to predict the z-score. The results are shown in Table 3. As all R^2 values are relatively small, thus no single or pair of branches makes a major contribution to the final analytic outcome. In other words, the outcome is highly variable and depends on many choices simultaneously rather than on just one or two choices.

The second method for quantifying branch sensitivity utilizes the *k*-samples Anderson Darling test (Scholz & Stephens, 1987). The *k*-samples Anderson Darling test measures the distance between the empirical distribution functions of *k* individual samples and that of the pooled sample. As each analytic approach has its own z-score distribution, the test quantifies how different these distributions are. Table 4 shows the standardized test statistics, with higher scores indicating more sensitive branches. In Figure 8, darker colors indicate more sensitive branches. For H1, DV and IV operationalizations lead to the most varied distributions in z-score, and for H2, alternative IV operationalizations have the most differing z-score distributions. However, the variance in estimates we were able to explain was again modest overall. Further details on the Boba multiverse are provided in Supplement 11.

5. Discussion

This crowdsourced investigation reveals striking dispersion in empirical results when many scientists address the same research question with the same data. When independent analysts tested two specific research predictions regarding the roles of gender and status in group meetings, they employed a wide array of approaches, which in turn led to a broad range of results. In a departure from previous many analyst projects, both variable operationalizations (e.g., how status is measured) and statistical analyses (e.g., covariate choices) were left unconstrained, contributing to the radical dispersion of estimates across independent analysts. Although the total variance in estimates we were able to explain was only modest, a Boba multiverse analysis (Liu et al., 2020) did demonstrate that variable operationalizations contributed most to radical dispersion in estimates, with statistical choices also contributing.

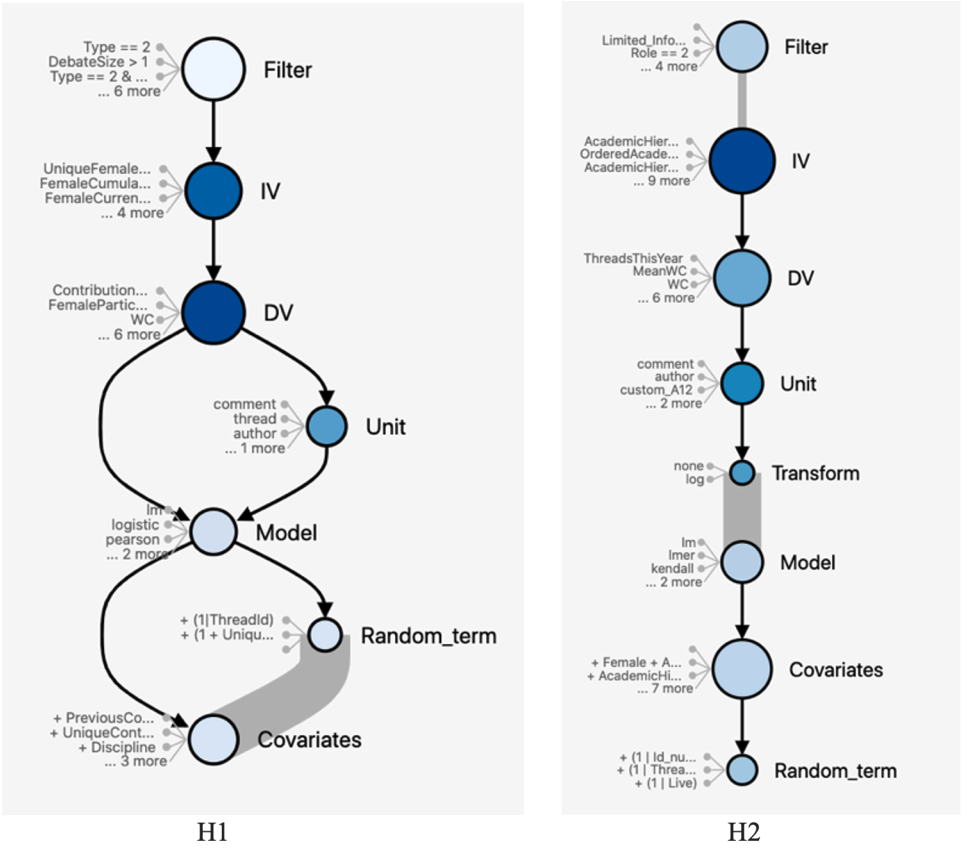


Fig. 8. Analytic decision graphs. Nodes represent analytic branches, and edges indicate order and dependency between branches. The size of a node encodes the number of alternative analytic approaches. Color maps to sensitivity, with darker color indicating a more sensitive branch. Here, sensitivity is computed using the k-samples Anderson-Darling test.

In Silberzahn et al. (2018) 69% of teams reported a statistically significant effect size in the expected direction, and no team reported a statistically significant effect size in the opposite of the predicted direction. In contrast, in the present initiative 64% of teams reported significant support for H1 and 29% for H2, with 21% and 21% reporting significant reversals, respectively. Such sign reversals are particularly strong evidence that subjective researcher choices make a critical contribution to the results obtained. This occurred under conditions closer to the typical research project, in which investigators must decide how to conceptualize and operationalize variables in addition to making statistical choices. The present pattern of results, which we term radical effect size dispersion, has never been demonstrated before in naturally occurring analyses by independent scientists. Situating the present findings, Table 5.1 describes projects that have crowdsourced various stages of the research process, and Table 5.2 summarizes the results of the crowdsourcing data analysis projects to date (see also Uhlmann et al., 2019). The present project is most similar in approach and results not to Silberzahn et al. (2018), but to Landy et al. (2020), who observed sign reversals across different experiments created by independent laboratories to test the same research question (i.e., conceptual replication designs, with operationalizations unconstrained).

Another key contribution of the present research is introducing and making publicly available the DataExplained tool developed for the project. Using the DataExplained portal, each participating researcher provided step-by-step explanations for her or his analytic decisions. Qualitative analyses of these reasonings about quantitative decisions led to the model of iterative research decision making shown in Figure 6. The DataExplained website (https://dataexplained.net/) is available for researchers who wish to carefully document their analytic decisions and justifications for them, either individually or as a crowd (see also the code in Supplement 9 and video demonstration at https://goo.gl/rnpgae). It is our hope that such platforms become a part of the organizational scholar's toolkit (Perkel, 2018) for transparently documenting her workflow. In the future, scientific journals like *Organizational Behavior and Human Decision Processes*, the *Journal of Management,* and the *Journal of Applied Psychology* may ask researchers to submit detailed documentation of their analytic steps and the reasons for the paths chosen (Aguinis et al., in press; Köhler et al., in press; Gelman & Loken, 2014), so that reviewers and readers can be convinced (or not) of the approach, and more easily formulate and run alternative specifications.

5.1. *Empirically explaining variability in results*

We were able to conclude from the Boba multiverse results (Liu et al., 2020) that how you think about your constructs (and thus operationalize variables) makes a contribution to radical dispersion in estimates (McGuire, 1973, 1983), in addition to statistical choices such as covariates and what type of regression or other measure of association is employed. For Hypothesis 1, dependent variable and independent variable operationalizations make the relatively largest contribution to dispersion in estimates across the multiverse of analytic approaches, and for Hypothesis 2 independent variable operationalizations was the single largest contributor. This highlights another level of subjectivity and researcher choice, in addition to the statistical choices previously examined by for example Silberzahn et al. (2018).

Although IV and DV choices do matter, their effect is small. Surprisingly (at least to us), the outcome of the analysis was only weakly related to one analytic choice or a combination of analytic choices. There are several possible causes for this unpredictability of the outcome

Table 3
The sensitivity of the branches according to adjusted R^2. Each cell represents the adjusted R^2 of one branch (diagonal) or the combination of two branches.

(a) *Hypothesis 1:*

	Filter	DV	IV	Covariates	Random term	Unit of analysis	Model
Filter	0.0007	0.0008	0.021	−0.009	0.002	−0.002	−0.004
DV	NA	0.007	0.038	0.003	0.007	0.006	0.007
IV	NA	NA	0.026	0.022	0.030	0.026	0.052
Covariates	NA	NA	NA	−0.0008	−0.0006	−0.0006	0.002
Random term	NA	NA	NA	NA	−0.0003	0.003	0.008
Unit of analysis	NA	NA	NA	NA	NA	0.001	0.002
Model	NA	NA	NA	NA	NA	NA	0.003

(b) *Hypothesis 2:*

	Filter	DV	IV	Random term	Covariates	Unit of analysis	Transform	Model
Filter	0.008	0.044	0.096	0.011	0.015	0.021	0.014	0.038
DV	NA	0.006	0.116	0.014	0.022	0.013	0.006	0.023
IV	NA	NA	0.045	0.055	0.062	0.073	0.069	0.063
Random term	NA	NA	NA	0.002	0.005	0.010	0.006	0.003
Covariates	NA	NA	NA	NA	0.004	0.014	0.012	0.005
Unit of analysis	NA	NA	NA	NA	NA	0.005	0.006	0.025
Transform	NA	NA	NA	NA	NA	NA	0.002	0.004
Model	NA	NA	NA	NA	NA	NA	NA	0.0004

of the analysis. First, the task for the analyst, testing a hypothesis with a dataset with yet unspecified variables, may have been so broadly formulated that the universe of potential analyses was enormous. This is confirmed by the actual analyses that differed in all respects; no two analyses were similar with respect to all analytic choices or the number of observations. Second, it may also be that the unpredictability of the outcome of the analysis reflects the nature of research in the social sciences; arbitrary choices may result in arbitrary outcomes of the analysis (see Figure 6). Of course, it is of paramount importance for social science research to distinguish the most important cause of diverging outcomes of multi-analyst projects. Does social science research have an intrinsically low rate of successful conceptual replications and reproducibility (Iso-Aloha, 2017; cf. Heino, Fried, & LeBel, 2017), or is it merely the characteristics of the current project that is responsible for the larger heterogeneity of results? Analyses of data of pre-registered many-lab studies suggest that minor changes to sample population and settings often do not affect the results and conclusions of experimental research (Olsson-Collentine, Wicherts, & van Assen, 2020). Hence, further crowdsourced data analysis initiatives testing many research hypotheses with many datasets, as well as further many-lab studies, are needed to address this question systematically.

5.2. More limitations and future directions

The analysts in the study were confronted with an unconventional research environment; a guiding theoretical framework was not directly provided by the project coordinators, and the dataset was sizeable with many variables that could potentially be used as operationalizations of the constructs in question (e.g., professional status). We therefore should be careful with generalizing the results to other research environments where, for instance, the theory is fully articulated and the dataset contains fewer variables and statistical choices to be made. Like any other research, crowd projects can and should be subjected to replication (Landy et al., 2020), and we believe a long series of crowdsourcing data analysis projects are necessary before drawing strong inferences from this line of research. At the same time, it is worth noting that the present dataset and the one leveraged by Silberzahn et al. (2018) are less complex than many archival datasets used by organizational scholars, economists, and others. With the present dataset, further operationalizations could have involved for example additional coding to quantify the amount of meaningful information conveyed per unit of text as a measure of verbal contributions to the debate. To the extent that complexity and ambiguity are positively correlated with dispersed results across different analysts, our findings may have wide implications for the conclusions drawn from analyses of complex datasets (see also Bastiaansen et al., 2020; Botvinik-Nezer et al., 2020; Silberzahn et al., 2018).

We fully acknowledge that our final crowd of analysts was relatively small (14 or 15 per hypothesis, for a total of 29 sets of analytic results), and heterogeneous in terms of job rank. Our results would perhaps be

Table 4
The sensitivity of the branches according to the k-samples Anderson Darling test. Each cell shows the standardized test statistics of a branch, with higher values indicating more sensitive branches.

(a) *Hypothesis 1:*

DV	IV	Unit	Model	Random terms	Covariates	Filter
275.79	238.85	160.07	49.55	42.88	39.80	12.26

(b) *Hypothesis 2:*

IV	Unit	Transform	DV	Random terms	Filter	Model	Covariates
421.08	294.83	253.39	229.46	156.92	134.36	125.69	121.57

Table 5.1
Crowdsourcing various stages of the research process with examples from the management and social psychology literatures.

Crowdsourced stage	Example	Description of approach	Outcome
Ideation	Schweinsberg et al. (present article)	A crowd of researchers was provided with a data descriptor and asked to nominate research questions for testing. A second crowd then voted on which hypotheses to test.	Crowd-generated hypotheses received independent ratings for scientific value as high as those generated by the project coordinators. Hypothesis 1 from the present article was crowd-generated.
Assembling resources	StudySwap	Online platform for posting research "needs" and "haves" (e.g., "needing" 200 participants from a particular nation or "having" a subject pool with participants of that nationality).	Laboratories successfully matched for replication projects and other collaborations (see https://osf.io/meetings/StudySwap).
Study design	Landy et al. (2020)	Up to 15 independent research teams designed brief online experiments testing up to 5 research questions.	For 4 out of 5 hypotheses, independent research teams designed experiments that returned significant estimates in the opposite direction from each other. Meta-analysing across the effect size estimates from the different designs, 2 of 5 hypotheses were robust across conceptual replications.
Data collection	Stewart, Chandler, and Paolacci (2017).	Online platforms such as Amazon's Mechanical Turk used to crowdsource data collections.	Large-sample data collections with lay adults greatly facilitated at low cost to the researchers.
Data analysis	Silberzahn et al. (2018)	Independent analysts test the same research question(s) using the same dataset.	Independent analysts use different specifications from one another and often obtain divergent results (see Table 5.2).
Writing research reports	Christensen and van Bever (2014)	Online platform used to collectively outline and draft a review article.	The article "The Capitalist's Dilemma" in Harvard Business Review.
Peer review	Open review	Peer review feedback from the submission process is published together with the final paper, and post-publication peer commentary is linked to the online version of the article.	Used for a subset of articles at the Open Psychology Journal (https://openpsychologyjournal.com/peer-review-workflow.php) and Meta-Psychology (https://open.lnu.se/index.php/metapsychology/; see also https://osf.io/3m4z3/) among others.
Replicating findings	Camerer et al. (2016)	A crowd of independent laboratories collect new data using the same experimental designs as in prominent published papers in experimental economics.	61% of selected findings from experimental economics successfully directly replicated (same method, new observations) by independent laboratories.
Deciding future directions	Lai et al. (2014, 2016)	Multi-round intervention contest aimed at optimizing interventions to reduce automatic associative preferences for White American relative to Black American targets.	Some research teams were able to improve the effectiveness of their intervention between rounds by observing the project results across interventions.

more convincing to some if more senior scholars, such as tenured faculty at highly ranked universities, were involved. The small final number of analysts is partly attributable to the scope of the task, specifically operationalizing and testing two hypotheses using a complex dataset while simultaneously explaining each decision taken (and not taken) using an online portal. The pool of potential analysts was further restricted to individuals well versed in R. The heterogeneity of seniority is a more general property of crowd research, which tends to attract interested parties from a diversity of career stages, something we see as a strength. Although our sample is far too small to draw strong inferences, an internal exploratory analysis suggests effect size dispersion in the present project was not driven by either more junior or more senior scientists (see Supplement 10). In Silberzahn et al. (2018), which featured a larger number of analysts ($N = 29$ teams), there was likewise no correlation between indices of seniority (e.g., job rank) and effect size estimates. Although the smaller sample in the present project facilitated carefully tracking of decisions with DataExplained as well as in-depth qualitative coding of each analysis (see more below), this came at the expense of running meaningful tests of the potential moderating roles of expertise and other analyst characteristics. Future projects with larger samples of analysts are needed to explore potential individual differences. To that end, Delios et al. (2020a) have recruited over 80 analysts to test four hypotheses from the field of strategic management using the same complex longitudinal dataset, assessing both statistical and topic expertise as potential moderators.

Further individual-differences that may shape researchers' choices should be investigated— for example, political beliefs may bias scientists towards analytic specifications that lead to ideologically consistent effect size estimates (Jelveh, Kogut, & Naidu, 2015). Although the present investigation and Silberzahn et al. (2018) examined gender and racial dynamics in group settings, Botvinik-Nezer et al. (2020) and Bastiaansen et al. (2020) observed variability in results across many researchers analyzing fMRI and event sampling data on non-politically charged topics, suggesting political biases are not necessary for dispersed effect size estimates to emerge across different investigators.

The specific hypothesis in question is also likely important, in that some research questions involve a greater number of theoretical frameworks and valid operationalizations of key variables. In the present initiative, Hypothesis 1 (Figure 3) was associated with comparatively more dispersed standardized scores than Hypothesis 2 (Figure 4). Although none of the hypotheses examined in the pilot exhibited convergence in results across analysts, there was still variability in the degree of divergence (Supplement 3). Thus, aspects of the research question may help explain dispersion in empirical results (see also Landy et al., 2020). There no doubt exists natural variability in the looseness of the construct-to-measure mapping across research questions. The difference is that in the standard, small-teams approach to science, one would typically never see the looseness, because the authors would usually only show the results for their chosen operationalizations.

Many scientific fields are currently worried about replicability, and archival researchers too have been increasingly concerned about both direct reproducibility (same data, same analysis) and robustness to different analytic approaches (same data, different analyses). The results of recent investigations suggest archival findings may be less robust than hoped when the same set of observations is used but a different analytic strategy is employed (Murphy & Aguinis, 2019; Orben & Przybylski, 2019; Silberzahn et al., 2018; Simonsohn et al., 2020; Steegen et al., 2016). It is also of interest to hold archival studies to the same replication standard to which experimental work is held—in other words, employing the same methodology and statistical analyses, but using new

Table 5.2
Overview of crowdsourcing data analysis projects to date.

	Description of dataset	Hypotheses or research question tested	Number of analysts	Degree of dispersion in results
Silberzahn et al. (2018)	Dataset of red card decisions across four major European football (soccer) leagues, with 146,028 referee-player dyads	Are soccer referees more likely to give red cards to dark-skin-toned players than to light-skin-toned players?	29 analysis teams	69% of analysis teams reported a statistically significant relationship such that light skin toned players received more red cards than dark skin toned players, whereas 31% did not. Estimates ranged from 0.89 to 2.93 in odds ratio units. No analysis team reported a statistically significant effect such that light skin toned players received relatively more red cards.
Botvinik-Nezer et al. (2020)	fMRI data from 108 research participants who performed a decision making task involving risk	Hypothesis 1: Positive parametric effect of gains in the vmPFC (equal indifference group) Hypothesis 2: Positive parametric effect of gains in the vmPFC (equal range group) Hypothesis 3: Positive parametric effect of gains in the ventral striatum (equal indifference group) Hypothesis 4: Positive parametric effect of gains in the ventral striatum (equal range group) Hypothesis 5: Negative parametric effect of losses in the vmPFC (equal indifference group) Hypothesis 6: Negative parametric effect of losses in the vmPFC (equal range group) Hypothesis 7: Positive parametric effect of losses in the amygdala (equal indifference group) Hypothesis 8: Positive parametric effect of losses in the amygdala (equal range group) Hypothesis 9: Greater positive response to losses in amygdala (equal range group vs. equal indifference group) Analysts were asked "whether each hypothesis was supported based on a whole-brain corrected analysis" (yes/no)	70 analysis teams	One of 9 hypotheses (H5) received statistically significant support across a large majority (84.3%) of teams. Three hypotheses were associated with nearly-uniform null results across analysts (94.3% non-significant findings). For the remaining five hypotheses between 21.4% and 37.1% of teams reported statistically significant support. At the same time, *meta*-analysis revealed significant convergence across analysis teams in terms of the activated brain regions they each identified.
Bastiaansen et al. (2020)	Experience sampling data from a single person	"What symptom(s) would you advise the treating clinician to target subsequent treatment on, based on a person-centered (-specific) analysis of this particular patient's ESM data?"	12 analysis teams	No team made similar recommendations regarding symptoms to target for treatment. The nature of identified symptoms varied widely. The 12 teams of independent analysts identified between 0 and 16 symptoms.
Schweinsberg et al. (present article)	Dataset on academic debates and their participants	Hypothesis 1: A woman's tendency to participate actively in a conversation correlates positively with the number of females in the discussion. Hypothesis 2: Higher status participants are more verbose than are lower status participants.	Up to 15 individual analysts per hypothesis	Different analysts reported statistically significant results in opposite directions for both Hypothesis 1 and Hypothesis 2 (see Table 2). Boba multiverse analysis demonstrates that variable operationalizations contribute to radical dispersion in estimates, above-and-beyond statistical choices.

observations. Delios et al. (2020b) are currently examining whether published findings from an ongoing stream of data on strategic management decisions generalize to other time periods and places. The reliability of archival findings is an important concern many scholars are working to address, both individually and in the context of crowd collaborations.

5.3. Potential solutions and countermeasures

The present results raise the possibility that many scientific findings reported by academic researchers, as well as statistical analyses by data scientists at firms and external consultancies, are not robust to different defensible operationalizations of variables and analytic choices. This sensitivity to investigator choices may remain unintentionally occluded under the traditional approach to research as conducted by individuals and small teams, in which relatively few analyses or approaches, often derived from a single theoretical and disciplinary perspective, are presented. Standard operating procedures and methodological path dependencies in an academic field or subfield may create an illusion of reliability, if other valid approaches are not attempted or included in research reports. Broadly consistent with the present findings, Landy et al. (2020) found that when up to 13 independent research teams designed their own experimental studies to address the same research question (e.g., "Are individuals who work in the absence of any material need to do so morally praised?"), the different study designs returned statistically significant effects in opposite directions for four out of five original ideas examined (see also Baribault et al., 2018). This converging evidence suggests that the link between subjective researcher choices and support for a given conclusion may be stronger than intuition suggests (see Botvinik-Nezer et al., 2020, who found that forecasters in a prediction market underestimated the impact of analytic choices on fMRI results).

The effort of a crowdsourced approach is most justified when dealing with controversial issues about which organizational scholars possess different prior beliefs (Leavitt, Mitchell, & Peterson, 2010), for research questions with important implications for public policies or organizational decision making, and for complex datasets in which a variety of defensible analytic approaches could be employed. Following the logic

of the wisdom of the crowds, in which aggregating estimates reduces individual level biases (Galton, 1907; Lorge, Fox, Davitz, & Brenner, 1958; Surowiecki, 2004), the central tendency of the effect size estimates calculated by many different analysts may provide a less subjective and error-prone estimate of the effect. For datasets that do not contain sensitive information, firms may consider websites like Upwork.com, Guru.com, StudySwap, Kaggle.com, and academic partners to help obtain independent perspectives. The aggregated results of a select crowd of statistical and topic experts might also be relied on (Mannes, Soll, & Larrick, 2014). However, aggregating different results is not completely justified when the estimated quantity differs radically from one set of analyses to the other. Further, even a strong consensus is no guarantee of validity, since consensus can result from shared (false) assumption— different analysts might operationalize status the same way due to shared values, or use the same easy-but-suboptimal statistical approach because they have all been trained the same way.

Although it has the benefit of creating transparency about the robustness of findings, recruiting a crowd of analysts is often inefficient and impractical (Uhlmann et al., 2019). Further, for many firms as well as organizational researchers, an important ethical limitation on crowdsourcing is confidentiality concerns (Aguinis et al., in press). Sensitive data, for example on a firm's employees, cannot be distributed to a dozen or more independent investigators so that their results can subsequently be compared. For the vast majority of cases in which crowdsourcing is not practical or ethical, individual researchers can employ multiverse analyses (Steegen et al., 2016) and specification curves (Simonsohn et al., 2020). The investigator generates as many defensible analytic strategies as she can, then carries out and reports numerous such specifications (see also Leamer, 1983, 1985; Muñoz & Young, 2018; Sala-i-Martin, 1997; Young & Holsteen, 2017), potentially leveraging the Boba multiverse approach to identify the most sensitive branches (Liu et al., 2020). Alternatively, a few external consultants and academic partners who have signed nondisclosure agreements, and data scientists within the firm might analyze the data independently of each other to see if their conclusions converge. For academics, another option is asking different researchers on the same team, or better yet members of an independent team, to separately conduct the analyses, then report both approaches in the article. Whether conducted individually, as independent copilots, or as a crowd, data analysis decisions should be rendered explicitly (e.g., using carefully commentated code, or the DataExplained platform at (https://dataexplained.net/) which can also be recreated and modified using the code provided in Supplement 9).

This study and other *meta*-scientific investigations into the robustness of research methodologies and results (Banks et al., 2016; Bedeian et al., 2010; Begley & Ellis, 2012; Bergh et al., 2017; Camerer et al., 2016, 2018; Chang & Li, in press; Ebersole et al., 2016; Klein et al., 2014; 2018; Landy et al., 2020; O'Boyle et al., 2019; Open Science Collaboration, 2015; Prinz, Schlange, & Asadullah, 2011) highlight the value of humility in communicating research findings, and caution in applying them in organizational decision making contexts. Each investigator interprets the data through her own lens and this is not only unavoidable, but perhaps even to be embraced. By leveraging the distributed knowledge, perspectives, and assumptions of diverse investigators, the true consistency of support for an empirical claim can be revealed.

Acknowledgements

The project was funded by a research grant from INSEAD and was also supported by the Swiss National Science Foundation under grant number 143411.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.obhdp.2021.02.003.

References

Aguinis, H., Banks, G.C., Rogelberg, S.G., Cascio, W.F. (in press). Actionable recommendations for narrowing the science-practice gap in open science. *Organizational Behavior and Human Decision Processes*.

Aguinis, H., & Solarino, A. M. (in press). Transparency and replicability in qualitative research: The case of interviews with elite informants. *Strategic Management Journal*. https://doi.org/10.1002/SMJ.3015.

Aguinis, H., Werner, S., Lanza Abbott, J., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods, 13*(3), 515–539. https://doi.org/10.1177/1094428109333339.

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. A. (2011). Public availability of published research data in high-impact journals. *PLoS One, 6*(9), Article e24357. https://doi.org/10.1371/journal.pone.0024357.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43*, 666–678.

Bamberger, P. A. (2019). On the replicability of abductive research in management and organizations: Internal replication and its alternatives. *Academy of Management Discoveries, 5*(2), 103–108.

Banks, G. C., Field, J. G., Oswald, F. L., O'Boyle, E. H., Landis, R. S., Rupp, D. E., et al. (2019). Answers to 18 questions about open science practices. *Journal of Business and Psychology, 34*(3), 257–270. https://doi.org/10.1007/s10869- 018-9547-8.

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology, 31*(3), 323–338. https://doi.org/10.1007/s10869-016- 9456-7.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Van Ravenzwaaij, D., et al. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences, 115*(11), 2607–2612.

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., et al. (2020). 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*, Article 110211.

Bedeian, A. G., Taylor, S. G., & Miller, A. N. (2010). Management science on the credibility bubble: Cardinal sins and various misdemeanors. *Academy of Management Learning & Education, 9*(4), 715–725. https://doi.org/10.5465/amle.9.4.zqr715.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*, 531–533.

Bergh, D. D., Sharp, B. M., Aguinis, H., & Li, M. (2017). Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. *Strategic Organization, 15*, 423–436.

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's threat to organizational research: Evidence from primary and meta-analytic sources. *Personnel Psychology, 69*(3), 709–750.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature, 582*, 84–88.

Brescoll, V. L. (2011). Who takes the floor and why: Gender, power, and volubility in organizations. *Administrative Science Quarterly, 56*, 621–640.

Brunner, J., & Schimmack, U. (2018). *Estimating population mean power under conditions of heterogeneity and selection for significance*. Manuscript submitted for publication. Available at: http://www.utstat.toronto.edu/~brunner/papers/Zcurve2.2.pdf.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*, 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science. *Nature Human Behaviour, 2*, 637–644.

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research, 42*(3), 294–320.

Carp, J. (2012a). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage, 63*(1), 289–300. https://doi.org/10.1016/j.neuroimage.2012.07.004.

Carp, J. (2012b). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience, 6*, 149. https://doi.org/10.3389/fnins.2012.00149.

Chang, A. C., & Li, P. (in press). Is economics research replicable? Sixty published papers from thirteen journals say "usually not." *Critical Finance Review*. http://dx.doi.org/10.17016/FEDS.2015.083.

Childers, C. P., & Maggard-Gibbons, M. (2020). Same data, opposite results?: A call to improve surgical database research. *JAMA Surgery*. https://doi.org/10.1001/jamasurg.2020.4991.

Christensen, C. M., & van Bever, D. (2014). The capitalist's dilemma. *Harvard Business Review, 92*, 60–68.

Cortina, J. M., Green, J. P., Keeler, K. R., & Vandenberg, R. J. (2017). Degrees of freedom in SEM: Are we testing the models that we claim to test? *Organizational Research Methods, 20*(3), 350–378. https://doi.org/10.1177/1094428116676345.

Delios, A., et al. (2020a). *Crowdsourcing data analysis 3*. Research project in progress.

Delios, A., et al. (2020b). *Can you step into the same river twice? Examining the context sensitivity of research findings from archival data*. Manuscript in preparation.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., et al. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68–82.

Feldman, M. (2018). *Crowdsourcing data analysis: Empowering non-experts to conduct data analysis*. Unpublished dissertation. University of Zurich.
Galton, F. (1907). Vox populi. *Nature, 75*, 7.
Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*, 460–465.
Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management, 41*(2), 632–643.
Grolemund, G., & Wickham, H. (2014). A cognitive interpretation of data analysis. *International Statistical Review, 82*(2), 184–204.
Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., et al. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Open Science, 5*(8), Article 180448. https://doi.org/10.1098/rsos.180448.
Heino, M. T., Fried, E. I., & LeBel, E. P. (2017). Commentary: Reproducibility in psychological science: When do psychological phenomena exist? *Frontiers in Psychology, 8*, 1004.
Hruschka, D. J., Schwartz, D., St, Cobb, John, D. C., Picone-Decaro, E., Jenkins, R. A., et al. (2004). Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field Methods, 16*(3), 307–331.
Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science, 11*, 365–371.
Iso-Ahola, S. E. (2017). Reproducibility in psychological science: When do psychological phenomena exist? *Frontiers in Psychology, 8. Article, 879*. https://doi.org/10.3389/fpsyg.2017.00879.
Jelveh, Z., Kogut, B., & Naidu, S. (2015). *Political language in economics*. Unpublished manuscript. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2535453.
Klein, G., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent systems, 21*(5), 88–92.
Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142–152.
Klein, R. A., Vianello, M., Hasselman, F., & Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490.
Köhler, T., Gonzàlez-Morales, M. G., Banks, G. C., O'Boyle, E., Allen, J., Sinha, R., Woo, S. E., & Gulick, L. (in press). Supporting robust, rigorous, and reliable reviewing as the cornerstone of our profession: Introducing a competency model for peer review. *Industrial and Organizational Psychology: Perspectives on Science and Practice*. https://doi.org/10.1017/iop.2019.121.
Kurasaki, K. S. (2000). Intercoder reliability for validating conclusions drawn from open-ended interview data. *Field Methods, 12*(3), 179–194.
Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., JoyGaba, J. A., et al. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*, 1765–1785.
Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., et al. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General, 145*, 1001–1016.
Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Ebersole, C. R., et al. (2020). Crowdsourcing hypothesis tests. *Psychological Bulletin, 146*(5), 451–479.
Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., et al. (2009). Computational social science. *Science, 323*(5915), 721–723.
Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review, 73*(1), 31–43.
Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review, 75*, 308–313.
Leavitt, K., Mitchell, T., & Peterson, J. (2010). Theory pruning: Strategies for reducing our dense theoretical landscape. *Organizational Research Methods, 13*, 644–667.
Liu, Y., Kale, A., Althoff, T., & Heer, J. (2020). Boba: Authoring and Visualizing Multiverse Analyses. *IEEE Transactions on Visualization and Computer Graphics (Proc. VAST)*.
Lorge, I., Fox, D., Davitz, J., & Brenner, M. (1958). A survey of studies contrasting the quality of group performance and individual performance, 1920–1957. *Psychological Bulletin, 55*, 337–372.
Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276–299.
MacCoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature, 526*(7572), 187–189.
McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2006). Lessons from the JMCB archive. *Journal of Money, Credit and Banking, 38*(4), 1093–1107.
McGuire, W. J. (1973). The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology, 26*(3), 446–456.
McGuire, W. J. (1983). A contextualist theory of knowledge: Its implications for innovations and reform in psychological research. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 16, pp. 1–47). New York, NY: Academic Press.
Muñoz, J., & Young, C. (2018). We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology, 48*(1), 1–33.
Murphy, K. R., & Aguinis, H. (2019). HARKing: How badly can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology, 34*(1), 1–17.
O'Boyle, E., Banks, G. C., Carter, K., Walter, S., & Yuan, Z. (2019). A 20-year review of outcome reporting bias in moderated multiple regression. *Journal of Business and Psychology, 34*(1), 19–37. https://doi.org/10.1007/s10869-018-9539-8.
O'Boyle, E. H., Jr, Banks, G. C., & Gonzalez-Mulé, E. (2017). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management, 43*(2), 376–399.
Obels, P.,Lakens, D., Coles, N.A.,Gottfried, J., & Green, S.A. (in press). Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science*.
Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin, 146*(10), 922.
Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). https://doi.org/10.1126/science.aac4716.
Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour, 3*, 173–182.
Paglieri, F. (2004). Data-oriented belief revision: Towards a unified theory of epistemic processing. In Onaindia & Staab, *Proceedings of STAIRS* (pp. 179-190). Amsterdam: IOS Press.
Patel, C. J., Burford, B., & Ioannidis, J. P. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology, 68*(9), 1046–1058.
Perkel, J. M. (2018). Open framework tackles backwards science. *Nature*. Available at: https://www.natureindex.com/news-blog/open-framework-tacklesbackwards-science.
Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews. Drug Discovery, 10*(9), 712.
R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org.
Rohrer, J., et al. (in press). Putting the self in self-correction: Findings from the Loss-of-Confidence Project. *Perspectives on Psychological Science*.
Sala-i-Martin, X. X. (1997). I just ran two million regressions. *The American Economic Review, 87*(2), 178–183.
Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE, 4*(9), Article e7078. https://doi.org/10.1371/journal.pone.0007078.
Saylors, R., & Trafimow, D. (in press). Why the increasing use of complex causal models is a problem: On the danger sophisticated theoretical narratives pose to truth. *Organizational Research Methods*. https://doi.org/10.1177/1094428119893452.
Schmid Mast, M. (2001). Gender differences and similarities in dominance hierarchies in same-gender groups based on speaking time. *Sex Roles, 34*, 547–556.
Schmid Mast, M. (2002). Dominance as expressed and inferred through speaking time: A meta-analysis. *Human Communication Research, 28*, 420–450.
Scholz, F. W., & Stephens, M. A. (1987). K-sample Anderson Darling tests. *Journal of the American Statistical Association, 82*(399), 918–924. https://doi.org/10.2307/2288805.
Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: *Undisclosed* flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.
Smerdon, D., Hu, H., McLennan, A., von Hippel, W., & Albrecht, S. (2020). Female chess players show typical stereotype-threat effects: Commentary on Stafford. *Psychological Science, 31*(6). https://doi.org/10.1177/0956797620924051.
Staub, N. (2017). *Revealing the inherent variability in data analysis*. Unpublished master's thesis, University of Zurich. https://doi.org/10.13140/RG.2.2.25745.53609.
Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences, 21*, 736–748.
Stockemer, D., Koehler, S., & Lentz, T. (2018). Data Access, transparency, and replication: New insights from the political behavior literature. *PS: Political Science & Politics, 51*(4), 799–803. https://doi.org/10.1017/S1049096518000926.
Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences, 115*(11), 2584–2589. https://doi.org/10.1073/pnas.1708290115.
Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York: Doubleday Books.
Silberzahn, R., & Uhlmann, E. L. (2015). Many hands make tight work: Crowdsourcing research can balance discussions, validate findings and better inform policy. *Nature, 526*, 189–191.
Silberzahn, R., Uhlmann, E. L., Martin, D., Anselmi, P., Aust, F., Awtrey, E., et al. (2018). Crowdsourcing data analysis: Do soccer referees give more red cards to dark skin toned players? *Advances in Methods and Practices in Psychological Science, 1*, 337–356.
Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour, 4*, 1208–1214. https://doi.org/10.1038/s41562-020-0912-z.
Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology, 67*(1), 415–437.
Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712.
Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation, 27*(2), 237–246.
Tukey, J. W., & Wilk, M. B. (1966). Data analysis and statistics: an expository overview. In *Proceedings of the November 7-10, 1966, fall joint computer conference* (pp. 695-709). Association for Computing Machinery.
Uhlmann, E. L., Ebersole, C., Chartier, C., Errington, T., Kidwell, M., Lai, C. K., et al. (2019). Scientific Utopia III: Crowdsourcing Science. *Perspectives on Psychological Science, 14*, 711–733.

Van't Veer, A., & Giner-Sorolla, R. (2016). Pre-registration in social psychology: A discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2–12.

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., et al. (2013). The availability of research data declines rapidly with article age. *Current Biology, 24*, 94–97. https://doi.org/10.1016/j.cub.2013.11.014.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632–638.

Weick, K., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science, 16*(4), 409–421.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*, 726–728.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832.

Williams, L. J., O'Boyle, E. H., & Yu, J. (2020). Condition 9 and 10 tests of model confirmation: A review of James, Mulaik, and Brett (1982) and contemporary alternatives. *Organizational Research Methods, 23*(1), 6–29. https://doi.org/10.1177/1094428117736137.

Womack, R. P. (2015). Research data in core journals in biology, chemistry, mathematics, and physics. *PLoS ONE, 10*(12), Article e0143460. https://doi.org/10.1371/journal.pone.0143460.

Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research, 46*(1), 3–40.

Young, C., & Horvath, A. (2015). Sociologists need to be better at replication. Retrieved at: https://orgtheory.wordpress.com/2015/08/11/sociologists-need-to-be-better-at-replication-a-guest-post-by-cristobal-young/.