

Same data, different conclusions: Radical dispersion in empirical results

when independent analysts operationalize and test the same hypothesis

1	Martin	Schweinsberg	ESMT Berlin
2	Michael	Feldman	University of Zurich
3	Nicola	Staub	University of Zurich
4	Olmo R.	van den Akker	Tilburg University
5	Robbie C.M.	van Aert	Tilburg University
6	Marcel A.L.M.	van Assen	Tilburg University and Utrecht University
7	Yang	Liu	University of Washington
8	Tim	Althoff	University of Washington
9	Jeffrey	Heer	University of Washington
10	Alex	Kale	University of Washington
11	Zainab	Mohamed	ESMT Berlin and Indiana University
12	Hashem	Amireh	ESMT Berlin and Humboldt University Berlin
13	Vaishali Venkatesh	Prasad	ESMT Berlin
14	Abraham	Bernstein	University of Zurich
15	Emily	Robinson	
16	Kaisa	Snellman	INSEAD
17	S. Amy	Sommer	Marshall School of Business, University of Southern California
18	Sarah M.G.	Otner	Imperial College Business School
19	David	Robinson	
20	Nikhil	Madan	Indian School of Business
21	Raphael	Silberzahn	University of Sussex Business School
22	Pavel	Goldstein	School of Public Health, University of Haifa
23	Warren	Tierney	University of Limerick
24	Toshio	Murase	Waseda University
25	Benjamin	Mandl	Stockholm School of Economics
26	Domenico	Viganola	Stockholm School of Economics
27	Carolyn	Strobl	University of Zurich
28	Catherine B.C.	Schaumans	Independent researcher
29	Stijn	Kelchtermans	KU Leuven
30	Chan	Naseeb	IBM
31	S. Mason	Garrison	Wake Forest University
32	Tal	Yarkoni	University of Texas at Austin

33	C.S. Richard	Chan	Stony Brook University
34	Prestone	Adie	University of Nairobi
35	Paulius	Alaburda	
36	Casper	Albers	University of Groningen
37	Sara	Alspaugh	University of California, Berkeley
38	Jeff	Alstott	Massachusetts Institute of Technology
39	Andrew A.	Nelson	University of Kentucky
40	Eduardo	Ariño de la Rubia	California State University-Dominguez Hills
41	Adbi	Arzi	INSEAD
42	Štěpán	Bahník	The Prague College of Psychosocial Studies
43	Jason	Baik	
44	Laura Winther	Balling	Copenhagen Business School
45	Sachin	Banker	University of Utah
46	David AA	Baranger	University of Pittsburgh
47	Dale J.	Barr	University of Glasgow
48	Brenda	Barros-Rivera	Texas A&M University
49	Matt	Bauer	Illinois Institute of Technology
50	Enuh	Blaise	Eskisehir Osmangazi University
51	Lisa	Boelen	Imperial College London
52	Katerina	Bohle Carbonell	Northwestern University
53	Robert A	Briers	Edinburgh Napier University
54	Oliver	Burkhard	
55	Miguel-Angel	Canela	University of Navarra
56	Laura	Castrillo	
57	Timothy	Catlett	
58	Olivia	Chen	
59	Michael	Clark	University of Michigan
60	Brent	Cohn	
61	Alex	Coppock	Yale University
62	Natàlia	Cugueró-Escofet	Universitat Oberta de Catalunya
63	Paul G.	Curran	Michigan State University
64	Wilson	Cyrus-Lai	INSEAD
65	David	Dai	St. Michael's Hospital, University of Toronto
66	Giulio Valentino	Dalla Riva	Department of Mathematics and Statistics, University of Canterbury
67	Henrik	Danielsson	Linköping University
68	Rosaria	de F. S. M. Russo	Universidade Nove de Julho
69	Niko	de Silva	ESMT Berlin
70	Curdin	Derungs	Lucerne University of Applied Sciences and Arts
71	Frank	Dondelinger	Lancaster University

72	Carolina	Duarte de Souza	Universidade Federal de Santa Catarina
73	B. Tyson	Dube	
74	Marina	Dubova	Indiana University
75	Ben Mark	Dunn	University of Glasgow
76	Peter Adriaan	Edelsbrunner	ETH Zurich
77	Sara	Finley	Pacific Lutheran University
78	Nick	Fox	Rutgers University
79	Timo	Gnambs	Leibniz Institute for Educational Trajectories & Johannes Kepler University Linz
80	Yuanyuan	Gong	Okayama University
81	Erin	Grand	
82	Brandon	Greenawalt	University of Notre Dame
83	Dan	Han	
84	Paul H. P.	Hanel	University of Bath, University of Essex
85	Antony B.	Hong	INSEAD
86	David	Hood	
87	Justin	Hsueh	
88	Lilian	Huang	University of Chicago
89	Kent N.	Hui	School of Management, Xiamen University
90	Keith A.	Hultman	Elmhurst College
91	Azka	Javaid	Columbia University Medical Center
92	Lily Ji	Jiang	University of Washington & Indiana University
93	Jonathan	Jong	University of Oxford & Coventry University
94	Jash	Kamdar	
95	David	Kane	Harvard University
96	Gregor	Kappler	University of Vienna
97	Erikson	Kaszubowski	Universidade Federal de Santa Catarina
98	Christopher M.	Kavanagh	
99	Madian	Khabsa	
100	Bennett	Kleinberg	University College London
101	Jens	Kouros	
102	Heather	Krause	York University

103	Angelos-Miltiadis	Kryptos	Department of Health Psychology, KU Leuven & Department of Clinical Psychology, Utrecht University
104	Dejan	Lavbič	
105	Rui Ling	Lee	Nanyang Technological University
106	Timothy	Leffel	The University of Chicago
107	Wei Yang	Lim	University of Colorado, Colorado Springs
108	Silvia	Liverani	Queen Mary University of London
109	Bianca	Loh	INSEAD
110	Dorte	Lønsmann	University of Copenhagen
111	Jia Wei	Low	Singapore Management University
112	Alton	Lu	University of Washington
113	Kyle	MacDonald	McD Tech Labs
114	Christopher R.	Madan	School of Psychology, University of Nottingham
115	Lasse Hjorth	Madsen	Novo Nordisk
116	Christina	Maimone	Northwestern University
117	Alexandra	Mangold	
118	Adrienne	Marshall	University of Idaho
119	Helena Ester	Matskewich	University of Washington
120	Kimia	Mavon	Harvard University
121	Katherine L.	McLain	ESMT Berlin
122	Amelia A	McNamara	University of St Thomas
123	Mhairi	McNeill	
124	Ulf	Mertens	Heidelberg University
125	David	Miller	Northwestern University
126	Ben	Moore	University of Edinburgh
127	Andrew	Moore	
128	Eric	Nantz	Eli Lilly
129	Ziauddin	Nasrullah	ESMT Berlin
130	Valentina	Nejkovic	University of Nis
131	Colleen S	Nell	George Washington University
132	Andrew Arthur	Nelson	University of Kentucky
133	Gustav	Nilsonne	Karolinska Institutet and Stockholm University
134	Rory	Nolan	University of Oxford
135	Christopher E.	O'Brien	
136	Patrick	O'Neill	University of Maryland, Baltimore County
137	Kieran	O'Shea	University of Glasgow

138	Toto	Olita	The University of Western Australia
139	Jahna	Otterbacher	Open University of Cyprus
140	Diana	Palsetia	Northwestern University
141	Bianca	Pereira	
142	Ivan	Pozdniakov	National Research University Higher School of Economics
143	John	Protzko	University of California, Santa Barbara
144	Jean-Nicolas	Reyt	McGill University
145	Travis	Riddle	National Institutes of Health / National Institute of Mental Health
146	Amal (Akmal)	Ridhwan Omar Ali	The University of Sheffield
147	Ivan	Ropovik	Charles University, Faculty of Education, Institute for Research and Development of Education & University of Presov, Faculty of Education
148	Joshua M.	Rosenberg	University of Tennessee, Knoxville
149	Stephane	Rothen	
150	Michael	Schulte- Mecklenbeck	University of Bern & Max Planck Institute for Human Development
151	Nirek	Sharma	Washington University in St. Louis
152	Gordon	Shotwell	Dalhousie University
153	Martin	Skarzynski	
154	William	Stedden	
155	Victoria	Stodden	University of Illinois at Urbana-Champaign
156	Martin A.	Stoffel	Institute of Evolutionary Biology, University of Edinburgh
157	Scott	Stoltzman	Colorado State University
158	Subashini	Subbaiah	CSU
159	Rachael	Tatman	Rasa Technologies
160	Paul H.	Thibodeau	Oberlin College
161	Sabina	Tomkins	Stanford University
162	Ana	Valdivia	University of Granada
163	Gerrieke B.	Druijff-van de Woestijne	Radboud University Nijmegen
164	Laura	Viana	University of Hawaii
165	Florence	Villesèche	Copenhagen Business School
166	W. Duncan	Wadsworth	Microsoft & Rice University
167	Florian	Wanders	University of Amsterdam
168	Krista	Watts	
169	Jason D	Wells	Dartmouth College
170	Christopher E.	Whelpley	College of Charleston
171	Andy	Won	

172	Lawrence	Wu	University of California, Berkeley
173	Arthur	Yip	
174	Casey	Youngflesh	Department of Ecology and Evolutionary Biology, University of California, Los Angeles
175	Ju-Chi	Yu	The University of Texas at Dallas, School of Behavioral and Brain Sciences
176	Arash	Zandian	Division of Affinity Proteomics, Department of Protein Science, KTH Royal Institute of Technology & SciLifeLab
177	Leilei	Zhang	
178	Chava	Zibman	
179	Eric Luis	Uhlmann	INSEAD

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 34 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.

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.

Contact authors: Martin Schweinsberg (martin.schweinsberg@esmt.org), Michael Feldman (mfeldman@peakdata.ch), Abraham Bernstein (bernstein@ifi.uzh.ch), Eric Luis Uhlmann (eric.uhlmann@insead.edu)

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.

Keywords: crowdsourcing data analysis; scientific transparency; research reliability; scientific robustness; researcher degrees of freedom; analysis-contingent results

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

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; Saylor & 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 et al., 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 et al., 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 Consciousness 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.

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) were 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, Kale, Althoff, & Heer, 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.

Methods

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. 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.

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 et al., 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%).

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 (<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, sub-sequences 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.

[Insert Figures 1 and 2 about here]

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).

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 <https://osf.io/n5q3c/>). These error-checks involved a two-step process. First, three research assistants from The European School of Management and Technology (ESMT) sub-team of the crowd project 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, 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 sub-teams 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.

Results

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, 1.2, and 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).

[Insert Tables 1.1, 1.2, and 1.3 about here]

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 than 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.

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 $(X-X_{\text{mean}})*(Y-Y_{\text{mean}})$ divided by a constant equal to $(X-X_{\text{mean}})^2$). 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 χ^2 -statistic (similar shape but skewed to the right), and for the F -statistic but only when $df_1=1$ (this is the t) or when df_1 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 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 $< .001$. 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 $z <- x/((x-y)/qnorm(.975))$). 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.

[Insert Figures 3 and 4 and Table 2 about here]

Qualitative coding of quantitative analytic decisions

That cognitive processes play a key role in data analysis has been 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 R-codes analysts 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).

[Insert Figure 5 about here]

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 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 et al., 2013; Kurasaki, 2000; Hruschka et al., 2004; Krippendorff, 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 .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 hypothesis 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 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.

[Insert Figure 6 about here]

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, 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.

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 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.

[Insert Tables 3 and 4 and Figures 7 and 8 about here]

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.

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.

[Insert Tables 5.1 and 5.2 about here]

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 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 important 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, and 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.

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 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 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.

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 (Steege 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 commented 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.

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*. doi: 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>.
- Alasuutari, P. (2010). The rise and relevance of qualitative research. *International Journal of Social Research Methodology*, 13(2), 139-155.
- 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), e24357. doi: 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., &

- Rogelberg, S. G. (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., ... & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, 115(11), 2607-2612.
- Barnes, C. M., Dang, C., Leavitt, K., Guarana, C., & Uhlmann, E. L. (2018). Archival data in micro organizational research: A toolkit for moving to a broader set of topics. *Journal of Management*, 44, 1453-1478.
- 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, 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>.
- Byington, E. K., & Felps, W. (2017). Solutions to the credibility crisis in management science. *Academy of Management Learning & Education*, 16(1), 142–162.
- 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. doi: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. doi: 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*. doi: 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., ..., & Nosek, B. A. (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., Nilsson, 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), 180448. doi: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., Cobb St. John, D. C., Picone-Decaro, E., Jenkins, R. A., & Carey, J.W. (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., ..., & Nosek, B. A. (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., González-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., ... & Nosek, B. A. (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., ... & Nosek, B. A. (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 Jr, E. H., 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). DOI: 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, September 6). Open framework tackles backwards science. *Nature*. Available at: <https://www.natureindex.com/news-blog/open-framework-tackles-backwards-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): e7078. doi: 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. doi: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), 956797620924051-759. doi: 10.1177/0956797620924051
- Staub, N. (2017). *Revealing the inherent variability in data analysis*. Unpublished master's thesis, University of Zurich. DOI: 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. doi: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. doi: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*. Doubleday Books, New York.

- 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., ..., & Nosek, B. A. (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.
- Steege, 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., McCarthy, R., Riegleman, A., Silberzahn, R., & Nosek, B.A. (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., ..., Rennison, D. J. (2013). The availability of research data declines rapidly with article age. *Current Biology*, 24, 94-97. <http://dx.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., Augusteyn, 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. <http://dx.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): 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/>

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
<u>1</u>	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
<u>2</u>	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
<u>3</u>	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
<u>4</u>	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
<u>5</u>	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

<u>6</u>	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
<u>7</u>	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
<u>9</u>	linear regression	634	Comments	None	number of words in a female comment	cumulative proportion of female comments in each conversation	23.47 regression coefficient
<u>11</u>	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
<u>12</u>	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
<u>13</u>	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

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

<u>10</u>	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
<u>11</u>	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
<u>12</u>	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
<u>14</u>	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
<u>17</u>	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

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.

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
<i>H1: A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion</i>	64.3% (<i>n</i> = 9)	0% (<i>n</i> = 0)	14.2% (<i>n</i> = 2)	21.4% (<i>n</i> = 3)
<i>H2: Higher status participants are more verbose than lower status participants</i>	28.6% (<i>n</i> = 4)	21.4% (<i>n</i> = 3)	28.6% (<i>n</i> = 4)	21.4% (<i>n</i> = 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).

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

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 et al. (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.

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.

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)

Botvinik-Nezer
et al. (2020)

fMRI data
from 108
research
participants
who performed
a decision
making task
involving risk

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's (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	<p>Hypothesis 1: A woman's tendency to participate actively in a conversation correlates positively with the number of females in the discussion.</p> <p>Hypothesis 2: Higher status participants are more verbose than are lower status participants.</p>	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.

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

Edit block

Please give a name to the block: *

regressions with square root and log transformation

Please shortly explain what you did in this block: *

Ran same regression as before, but with log and square root transformations of predictors.

What were the other (if any) alternatives you considered in order to achieve the results of this block?

Please describe each alternative and explain its advantages and disadvantages. By clicking on "Add another alternative", you can add additional alternatives.

Alternative

No transformation of predictors

Advantages of this alternative

Better interpretability

Disadvantages of this alternative

Potential for slightly worse diagnostic plots (heteroscedasticity, skewness of residuals)

ADD ANOTHER ALTERNATIVE

Why did you choose your option? *

I experimented with both, but will ultimately use the non-transformed data for reporting; diagnostic plots did not improve much with

What preconditions should be fulfilled to successfully execute this block? *

previous data wrangling

SHOW DIFF

DELETE BLOCK

LOAD FILES

SAVE

CANCEL

```

fit3 <- lm(comments_now_percent_change ~
log(UniqueFemaleContributors),
data = reg_dat[-244,])
summary(fit3)
plot(fit3)
fit4 <- lm(comments_now_percent_change ~
sqrt(UniqueFemaleContributors),
data = reg_dat[-244,])
summary(fit4)
plot(fit4)

```

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

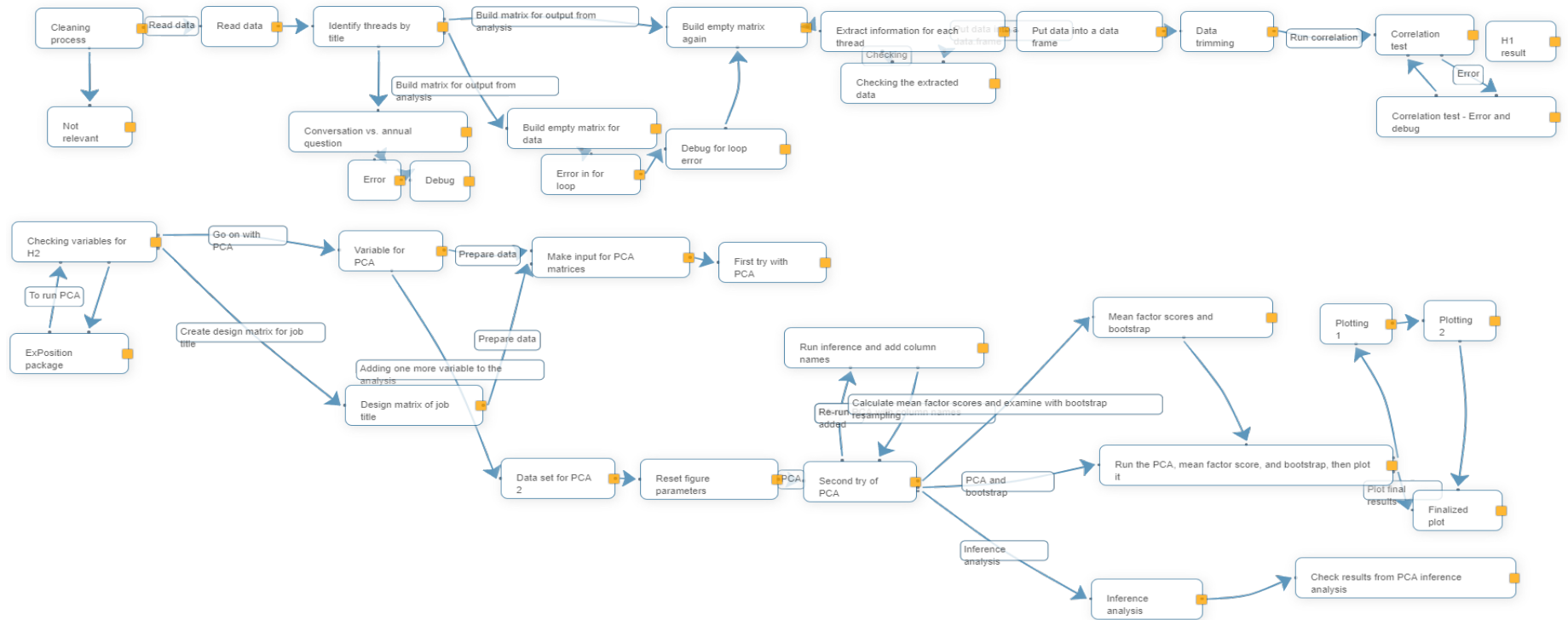
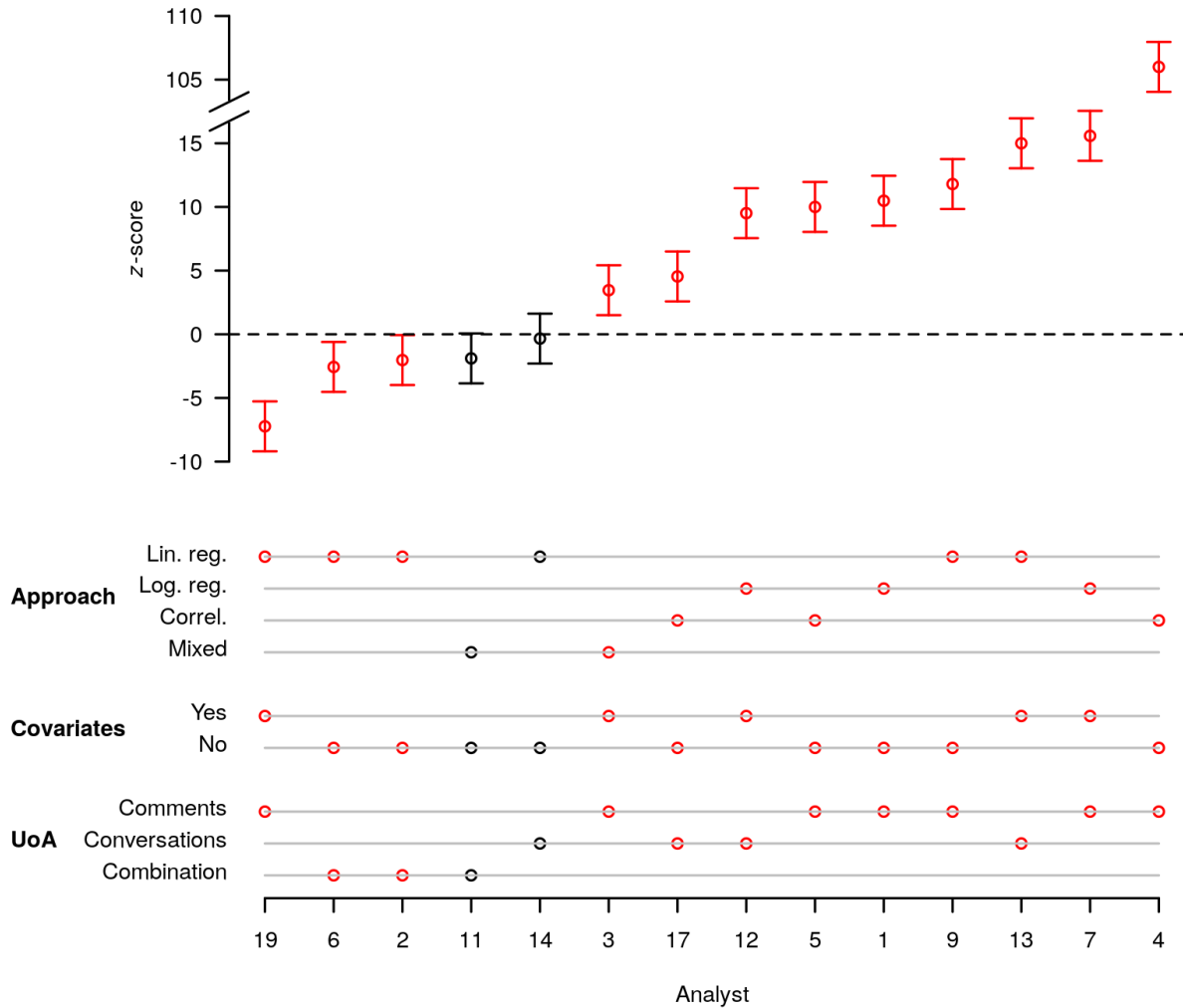


Figure 3. Dispersion of z-scores corresponding to estimates of independent analysts using the same dataset to test Hypothesis 1 (“A woman’s tendency to participate actively in the conversation correlates positively with the number of females in the discussion”), together with some details on each specification.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



Note. Analyst 12 used a Poisson regression model to test hypothesis 1 and this was categorized under logistic regression in the figure.

Figure 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.

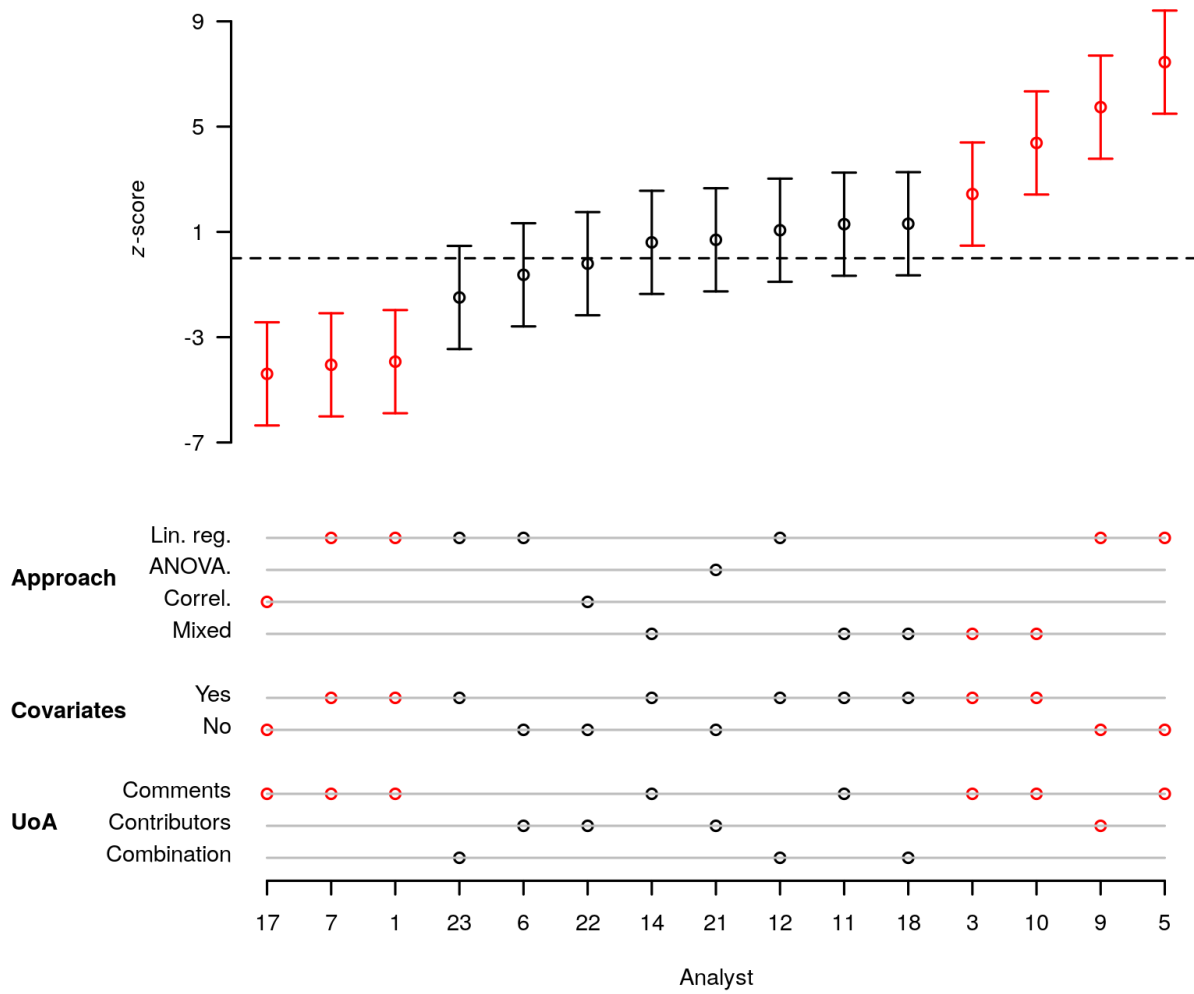


Figure 5. The workflow of our qualitative analysis of the quantitative analytic decisions.

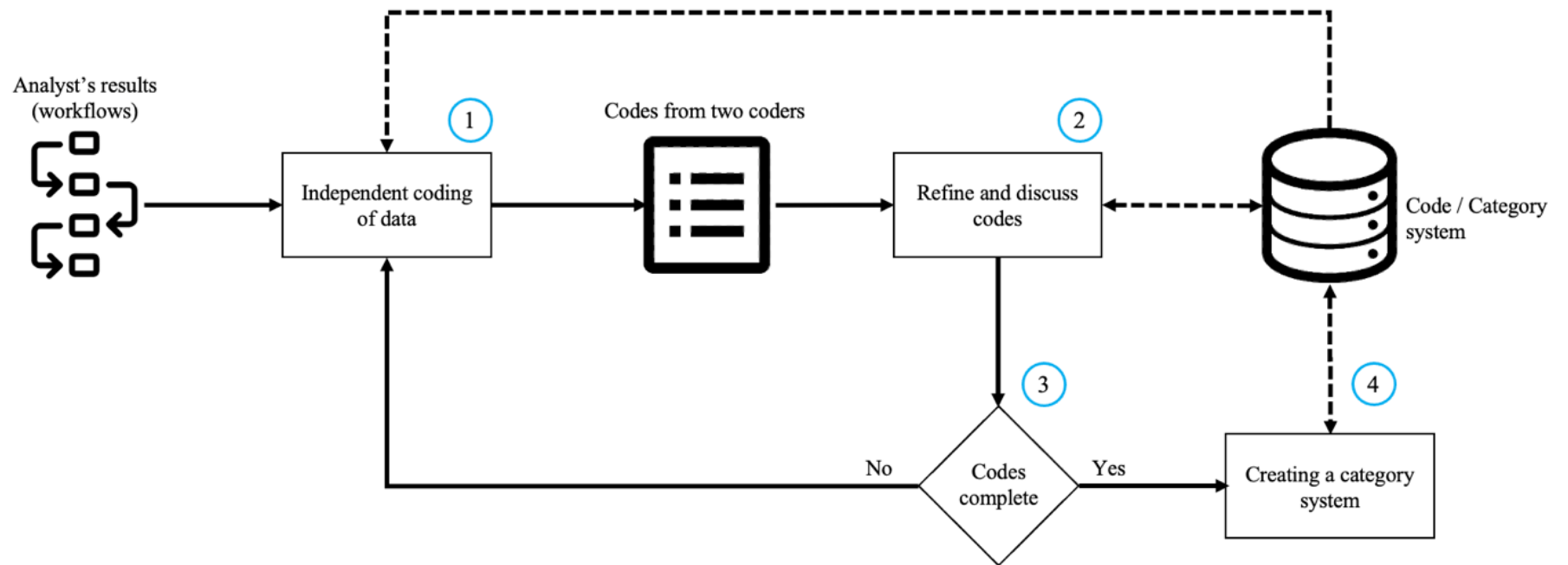


Figure 6. Model of an analyst's reasoning process.

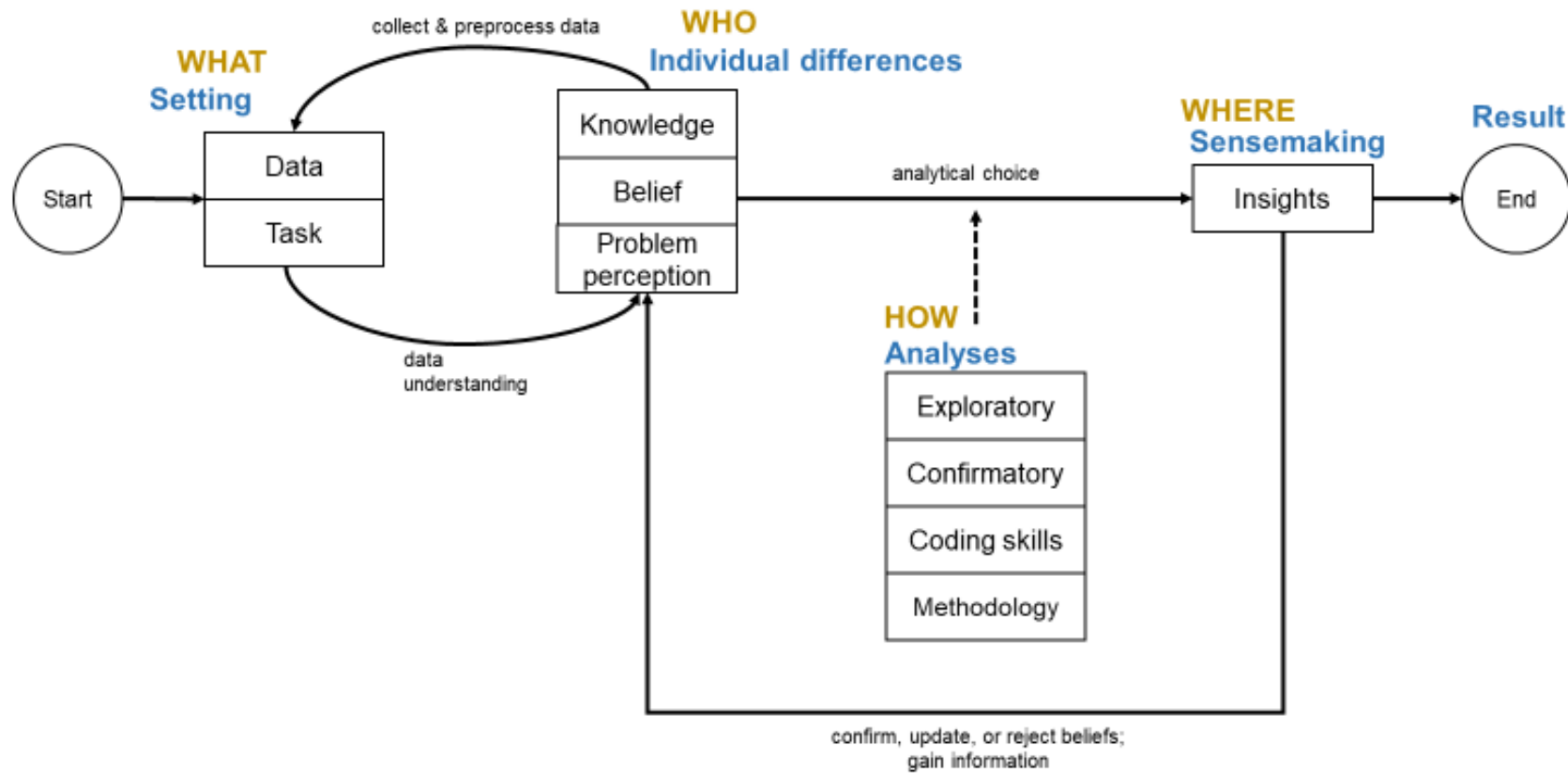
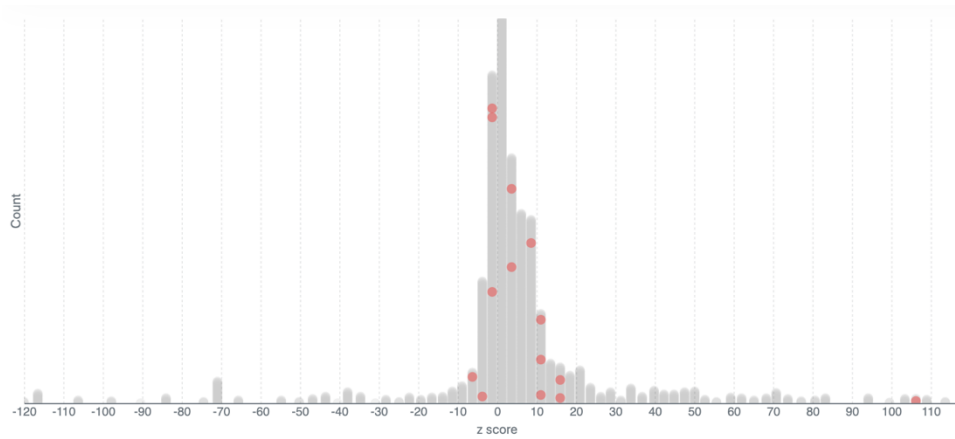


Figure 7. In the Boba multiverse analysis, z-scores for Hypotheses 1 and 2. Outcomes from the crowd analysts are highlighted in red and represent only a subset of the multiverse of possible analyses.

a) *z-scores for Hypothesis 1:*



b) *z-scores for Hypothesis 2:*

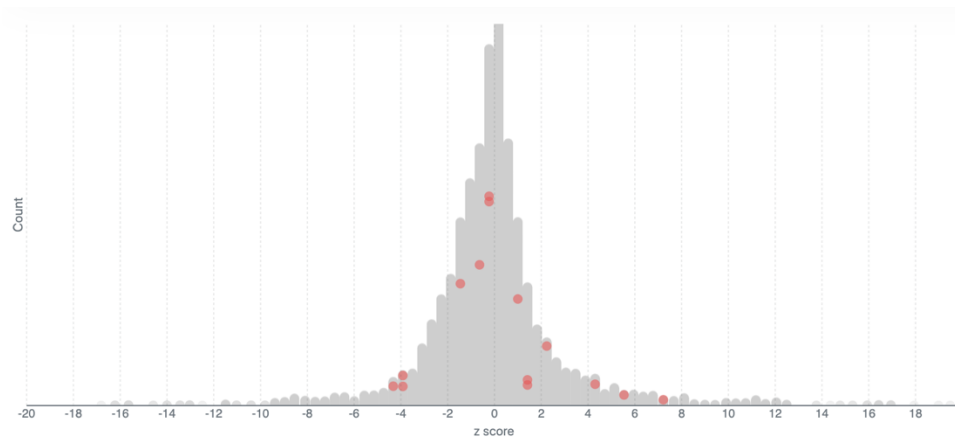
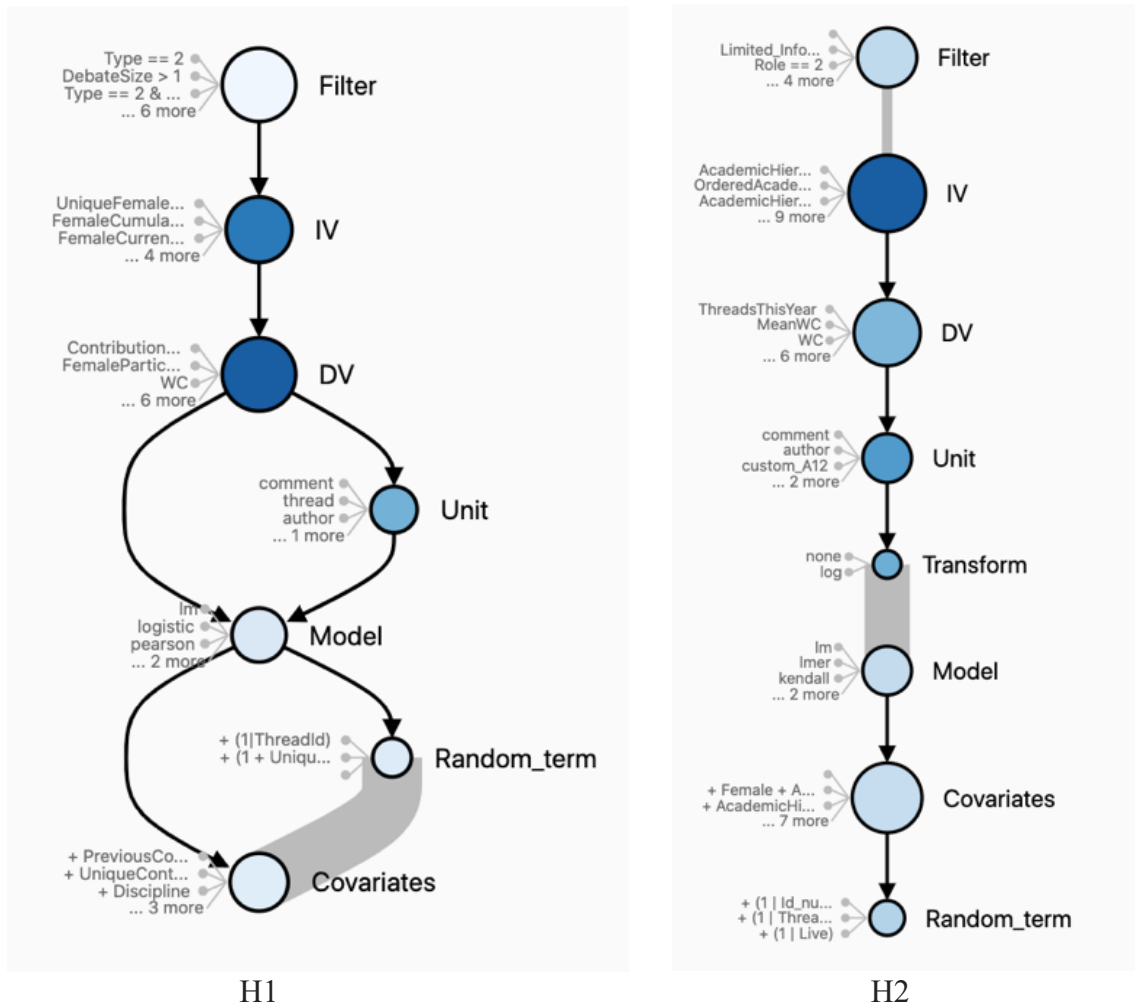


Figure 8. Analytic decision graphs. Nodes represent analytic branches, and edges indicate order and dependency between branches. The size of a node encodes the number of alternative analytic approaches. Color maps to sensitivity, with darker color indicating a more sensitive branch. Here, sensitivity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



**Supplements for “Same data, different conclusions: Radical dispersion in empirical results
when independent analysts operationalize and test the same hypothesis”**

Table of Contents:

<i>Supplement 1: Overview of intellectual debates dataset.....</i>	<i>3</i>
<i>Supplement 2: Crowdsourced hypothesis generation and selection</i>	<i>10</i>
<i>Supplement 3: Pilot study</i>	<i>25</i>
<i>Supplement 4: Online advertisements for primary project.....</i>	<i>41</i>
<i>Supplement 5: Pre-survey for analysts in primary study</i>	<i>44</i>
<i>Supplement 6: Post-survey for analysts in primary study</i>	<i>51</i>
<i>Supplement 7: Quality checks of crowdsourced data analyses.....</i>	<i>54</i>
<i>Supplement 8: Error checks of crowdsourced data analyses.....</i>	<i>60</i>
<i>Supplement 9: Qualitative analyses of explanations for analytic decisions.....</i>	<i>106</i>
<i>Supplement 10: Exploratory analysis on analyst expertise and effect size dispersion.....</i>	<i>155</i>
<i>Supplement 11: Further details on the Boba multiverse analyses.....</i>	<i>156</i>

Supplement 1: Overview of intellectual debates dataset

Our dataset build started with collecting information from Edge.org on all of the conversations and annual questions. We built a program that downloaded the information from the website, including the year, title, link to, and type of the conversation, as well as the text itself and who said it. Two independent coders then coded gender of the contributors based on their profile picture on Edge.org, or, if that was not available, pictures and pronouns on other reputable websites. Two research assistants, with the help of members of the coordination team then manually collected information on the job title, workplace, and PhD by finding CVs, university webpages, news articles, personal websites, and LinkedIn profiles. We wrote a program to collect the US News and World Report International Rankings and the Shanghai Rankings and manually gathered the rankings from the National US News and World Report Rankings. We then ran the text of the conversations and annual question responses through the LIWC program (Pennebaker, Francis, & Booth, 2003). Finally, we calculated the rest of the variables (such as number of female contributors to a conversation, previous contributions by the person to the Edge, etc.) based on the data we had already collected.

The descriptions below include the variable names in the full version of the dataset as well as the shortened variable name used in the dataset for older software. We indicate manually collected variables in the list below by including “(collected manually)” in the variable description.

- **Conversation level:**
 - **Year:** The year when the conversation took place.
 - **Title:** The title of the conversation. For example: “What scientific idea is ready for retirement?”
 - **Link:** A link to the conversation.
 - **Type:** 1 for annual question, 2 for conversation.
 - Edge does an **annual question** every year; some examples are “What scientific idea is ready for retirement?” and “What will change everything?” People then write in with their answers. All of the contributions are therefore written and asynchronous.
 - What Edge refers to as a **conversation** can actually be multiple things. Some of these are written essays by a single person, some are transcripts of a speech, and some are transcripts of a conversation (either between two or more guests or an interview).
- **ThreadID:** A unique identifier for each conversation/annual question (between two or more people).
- **Male_Contributions:** The number of times a man speaks in a specific conversation, it does not always equal the number of unique men in a conversation (see below).
- **Female_Contributions:** The number of times a woman speaks in a specific conversation, it does not always equal the number of unique women in a conversation (see below).
- **FemaleParticipation:** Female contributors/(number of total contributions); the percentage of comments that are made by a woman
- **NumberofAuthorContributions:**

- For the annual questions, this equals 0; because the website is the “author” of the question, everyone are considered commentators.
- Otherwise, this is the total number of times people contribute to the main body of the text, rather than people who just comment. For example, in <http://edge.org/conversation/how-democracy-works-or-why-perfect-elections-should-all-end-in-ties>, there are multiple people commenting on the post, but W. Daniel Hillis is the only author and only speaks once (as it is an essay). So NumberAuthors is “1.” If two people each spoke five times in a dialogue, NumberAuthors would be “10.”
- **DebateSize:** Number of text pieces in a conversation; this is the sum of female and male contributions.
- **Live:** Whether the text piece was transcribed or written; it is 0 if it is written (either an essay or a comment on a piece) and 1 if it was part of a live conversation or speech that was later transcribed. Here are the types of text and how they would be classified:
 - **A single author essay** (live = 0 because it is written): <http://edge.org/conversation/the-evolved-self-management-system>
 - **A single author speech** (live = 1 because it was spoken and later transcribed): <http://edge.org/conversation/cities-as-gardens>
 - **A live conversation**, either between multiple people or in an interview format (live = 1 because it was spoken and later transcribed): <http://edge.org/conversation/japan-inc-meets-the-digerati>
 - **Online Comments** on any of the three types above (live = 0 because it was written).
 - **The annual question (Type = 1):** live = 0 because these were all written and submitted.
- **UniqueContributors:** UniqueMaleContributors + UniqueFemaleContributors.
- **UniqueMaleContributors:** The number of unique male contributors.
- **UniqueFemaleContributors:** The number of unique female contributors.
- **UniqueFemaleParticipation:** The percentage of unique female participants; UniqueFemaleContributors divided by UniqueContributors.

Participant Level

- **Id:** The unique identifier of the contributor.
- **Id_num:** The unique identifier of the contributor as text (this is typically in the format of first name_last name).
- **Role:** Either author (=1) or commentator (=2).
- **TwoAuthors:** Some of the Edge comments are written by two people. In this case, we duplicated the row and kept the text level and conversation level information the same and had one author per row. This variable is 1 if this text was written by two people and 0 otherwise.
- **Name:** Name of the commentator [Anonymized in the publicly posted dataset].
- **Male:** (collected manually) The commentator is female = 0; the commentator is male = 1.
- **Female:** (collected manually) The commentator is male = 0, the commentator is female = 1.

- **Academic:** 1 = the person is in academia, 0 = they are not.
- **Limited_Information:** Equals 1 if we could only find limited information about the person (e.g. they commented in 2013 but we only have their job title from 2012), 0 otherwise.
- **Job_Title:** (collected manually) The job title of the commentator.
- **Job_Title_S:** (collected manually) This is a simplified list of job titles (e.g. we have “Eugene Higgs Professor” in Job.Title but “Chaired Professor” in Job.Title.Collapsed).
 - Chaired Professor
 - Professor
 - Associate Professor
 - Assistant Professor
 - Non-Tenure-Track Faculty
 - Postdoctoral Researcher
 - Graduate Student
 - Academic Leadership (Dean, Vice President, etc.)
 - Researcher
 - Artist/Author/Editor/Writer
 - Director
 - Founder
 - Other
 - Top Management and Founder
 - Top Management
 - Entrepreneur
 - Not Available
- **Job_Title_S_num:** (collected manually) Job_Title_S as numbers instead of text.
- **Department:** (collected manually) What academic department someone is in.
- **Department_S:** (collected manually) A simplified version of all the departments (e.g. while Jane Smith’s Department is “Experimental Physics,” her Department_S is “Physics”).
 - Physics (Phy)
 - Anthropology (Ant)
 - Earth Sciences (ES)
 - Biology (Bio)
 - Psychology (Psych)
 - Journalism, media studies and communication (JMS)
 - Medicine (Med)
 - Philosophy (Phil)
 - Space Sciences (SS)
 - Linguistics (Lin)
 - Computer Sciences (CS)
 - Engineering (Eng)
 - Arts (Arts)
 - Business/Management (Bus)
 - Environmental Studies and Forestry (ESF)
 - Sociology (Soc)

- Mathematics (Math)
- Asian Studies (AS)
- Education (Educ)
- Political Science (PS)
- Economics (Econ)
- Systems Science (Sys)
- History (Hist)
- Music (Musc)
- Chemistry (Chem)
- Archeology (Arch)
- Architecture and Design (ArchD)
- Law (Law)
- Zoology (Zoo)
- Literature (Lit)
- Divinity (Div)
- **Department_S_num:** (collected manually) Department_S as numbers instead of text.
- **Discipline:** This groups academic departments into disciplines.
 - Natural Sciences (NS)
 - Social Sciences (SocS)
 - Professions (Prof)
 - Humanities (Hum)
 - Formal Sciences (FS)
- **Workplace:** (collected manually) Where someone works; some people are self-employed
- **HavePhD:** (collected manually) Equals 1 if they have a PhD, 0 otherwise. It is 1 even if someone earns a PhD after they comment (e.g. John Doe comments in 2000 and earns his PhD in 2012; his comment in 2000 will still have HavePhD = 1)
- **PhD_Field:** (collected manually) What field contributors got their PhD in.
- **PhD_Year:** (collected manually) What year they got their PhD.
- **PreviousContributions:** How many times **before this year** this person has made contributions to the Edge. So if Jane Doe only talked three times in one conversation in 2012 and one time each in two conversations in 2014 (and never made any other comments), this will be 0 for her comment in 2012 and 3 for both her comments in 2014.
- **ContributionsThisYear:** How many times they contributed this year; even if they only participated in one conversation, if they spoke 40 times in that conversation, this variable will be 40.
- **ThreadsThisYear:** How many threads they participated in this year; thus if Melanie spoke in two threads in 2014, one twenty times and one once, this would equal 2 in 2014, while **ContributionsThisYear** would equal 21 for 2014.
- **PreviousThreads:** How many threads they participated in **before this year**. So if Lisa contributed for the first time twice in one thread in 2000, once each in two different threads in 2004, and once in 2014, this would be 0 for 2000, 1 for 2004,

and 3 for 2014 (and for **PreviousContributions** it would be 0 for 2000, 2 for 2004, and 4 for 2014).

- **AuthorandCommentator:** If, for the same piece, someone is both an author and a commentator, this is 1 for that person for that piece; otherwise it is 0.
- **PhD_Institution:** (collected manually) At what institution the commentator got their PhD.
- **Years_from_PhD:** (collected manually) How many years at the time of the comment since they earned their PhD; this is just Year - PhD.Year. This can be negative because people may have earned their PhD years after they make a comment.
- **PhD_Institution_SR:** The Shanghai Rankings of their PhD Institution; this is only for people who received their PhDs from institutions that are ranked by Shanghai. Shanghai ranks between 500 and 510 universities worldwide each year and also bins their rankings after a certain point, in different ways for different years (e.g. a university may be ranked as 301-352).
- **PhD_Institution_SR_Bin:**
 - 1 = university was ranked between 1 and 50
 - 2 = university was ranked between 51 and 100
 - 3 = university was ranked between 101 and 150
 - 4 = university was ranked between 151 and 200
 - 5 = university was ranked between 201 and 300
 - 6 = university was ranked between 301 and 400
 - 7 = university was ranked between 401 and 510
- **Workplace_SR:** The Shanghai Rankings of their workplace; this is only for academics and academic institutions that are ranked by Shanghai (see PhD_Institution_SR for more information).
- **Workplace_SR_Bin:**
 - 1 = university was ranked between 1 and 50
 - 2 = university was ranked between 51 and 100
 - 3 = university was ranked between 101 and 150
 - 4 = university was ranked between 151 and 200
 - 5 = university was ranked between 201 and 300
 - 6 = university was ranked between 301 and 400
 - 7 = university was ranked between 401 and 510
- **SR_Ranking_Dif:** The difference between the binned Shanghai Ranking University of their workplace and the binned Shanghai Ranking of their PhD; a positive ranking means that they work at a place that has a higher ranking than where they got their PhD.
- **PhD_Institution_US_IR:** (collected manually) The US News and World Report created an international ranking system in 2014 to rank the top 500 universities. Thus, even if a comment was made in 1999, if they have a PhD from Carnegie Mellon, this ranking will be Carnegie Mellon's ranking in the 2014 report.
- **PhD_Institution_US_IR_Bin:** (collected manually)
 - 1 = university was ranked between 1 and 50
 - 2 = university was ranked between 51 and 100
 - 3 = university was ranked between 101 and 150

- 4 = university was ranked between 151 and 200
- 5 = university was ranked between 201 and 250
- 6 = university was ranked between 251 and 300
- 7 = university was ranked between 301 and 350
- 8 = university was ranked between 351 and 400
- 9 = university was ranked between 401 and 450
- 10 = university was ranked between 451 and 500
- **Workplace_US_IR:** See PhD_Institution_US_IR.
- **Workplace_US_IR_Bin:**
 - 1 = university was ranked between 1 and 50
 - 2 = university was ranked between 51 and 100
 - 3 = university was ranked between 101 and 150
 - 4 = university was ranked between 151 and 200
 - 5 = university was ranked between 201 and 250
 - 6 = university was ranked between 251 and 300
 - 7 = university was ranked between 301 and 350
 - 8 = university was ranked between 351 and 400
 - 9 = university was ranked between 401 and 450
 - 10 = university was ranked between 451 and 500
- **USA_I_Ranking_Dif:** The difference between the rank of someone's workplace and the rank of their PhD Institution (as ranked by US News and World Report International Rankings). If this is positive, it means they're working at an institution ranked higher than their PhD Institution.
- **PhD_Institution_US:** The ranking of their PhD Institution by USA News and World Report; this is only for US institutions and only for a limited number of them. Different numbers of school were ranked in different years; for example, 129 schools were ranked in 2005, while only 51 were ranked in 2003. These numbers only span from 2003-2014.
- **PhD_Institution_US_Bin:**
 - 1 = university was ranked between 1-5
 - 2 = university was ranked between 6-10
 - 3 = university was ranked between 11-25
 - 4 = university was ranked between 26-50
 - 5 = university was ranked between 51-100
 - 6 = university was ranked between 101-150
 - 7 = university was ranked between 151-200
- **Workplace_US:** The ranking of their workplace by USA News and World Report; this is only for US institutions and only for a limited number of them. Different numbers of school were ranked in different years; for example, 129 schools were ranked in 2005, while only 51 were ranked in 2003. These only span from 2003-2014.
- **Workplace_US_Bin:**
 - 1 = university was ranked between 1-5
 - 2 = university was ranked between 6-10
 - 3 = university was ranked between 11-25
 - 4 = university was ranked between 26-50

- 5 = university was ranked between 51-100
- 6 = university was ranked between 101-150
- 7 = university was ranked between 151-200
- **USA_Ranking_Dif:** The difference between the rank of someone's workplace and the rank of their PhD Institution (as ranked by US News and World Report Rankings). If this is positive, it means they're working at an institution ranked higher than their PhD Institution.
- **Total_Citations:** The total number of citations they have received, including that year and all previous years (it's citations.year + previous citations).
- **H_Index:** This is their h-index in **2014**; a scholar has an index of h if they have published h papers each of which has been cited in other papers at least h times.
- **i10_Index:** How many papers in **2014** they had authored that has more than 10 citations; this is only for contributors with Google Scholar pages. As the Google Scholar pages only have an i10 index from 2014, even if the comment was from 1999, the i10 index is from 2014.
- **Citations_Year:** How many citations they received this year; this is only for contributors with Google Scholar pages.
- **Citations_Cumulative:** How many citations they have received in this year and previous years; this is only for contributors with Google Scholar pages
- **AcademicHierarchyStrict:**
 - 1 = Graduate Student
 - 2 = Postdoctoral
 - 3 = Assistant Professor
 - 4 = Associate Professor
 - 5 = Professor
 - 6 = Chaired Professor
- **PreviousCitations:** The number of citations they have received in all of the previous years
- **ContributionsbyAuthor:** The number of contributions by this author in this conversation
- **Dummy variables for Discipline (36 total)**

Text-Level

- **Order:** The order of the text pieces. Note this is meaningless for Annual Questions.
- **Text:** The text of the conversation.
- **LIWC variables:** See <https://www.liwc.net/LIWC2007LanguageManual.pdf>
- **Number.Characters:** Number of characters in the text piece

Reference for Supplement 1

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2003). Linguistic inquiry and word count: LIWC2001 manual. Mahwah, NJ: Erlbaum.

Supplement 2: Crowdsourced hypothesis generation and selection

The specific hypotheses to be tested using the “gender, status, and science” dataset were generated in two ways. First, the project coordination team developed a number of hypotheses based on their reading of the existing literature and theory. These included: high status scientists speak more and use more dominant language than low status scientists; male scientists speak more and use more dominant language than female scientists; high status is a better predictor of verbosity and dominant language for male scientists than female scientists; and very low and very high status participants are the least likely to speak at length and use dominant language.

Second, a crowdsourced approach was used to generate further research ideas. A crowd of 78 scientists who had expressed interest in the project based on social media advertisements (on Twitter and Facebook) were provided information about the dataset structure and variables, and asked to nominate hypotheses anonymously. A total of ten research ideas were proposed by members of the crowd. In a follow-up survey, the same crowd of scientists were presented with each hypothesis and asked to provide numeric ratings for its likelihood of being true (*1 = very unlikely to 7 = very likely*), interest value if true (*1 = not at all scientifically interesting to 5 = extremely scientifically interesting*), and for their overall assessment of whether it should be subjected to a systematic empirical test (*1 = strongly disagree to 7 = strongly agree*) (see Appendices S2-1 and S2-2 for the complete surveys and <https://osf.io/xr38c/> for the data).

Table S2-1 presents the ten collectively generated hypotheses along with the aggregated evaluations of their scientific value. Several of these crowd-generated hypotheses (marked with an asterisk) were selected for inclusion in the pilot analysis phase based on their “whether to test” ratings. Some favorably rated hypotheses were excluded because they either required adding new variables to the dataset (e.g., “Male contributors will be less likely to refer to the work of others in their own responses”) or called for data-analytic techniques many analysts would not be familiar with (e.g., “High status contributors coordinate their linguistic style less than low status contributors”).

Table S2-1: *Crowdsourced hypotheses and their evaluations on dimensions of scientific value*

Crowdsourced hypotheses	Interest value if true	Likelihood of being true	Vote whether to test hypothesis
<i>*Male commenters will comment more frequently and at greater length than female commenters. This will be especially true for live conversations, and attenuated/nonexistent for the non conversation format comments.</i>	2.80	4.92	5.38
<i>The main effect of gender on dominant language and verbosity would be higher in the academic subpopulation than in the nonacademic cluster.</i>	2.97	4.06	4.45
<i>The main effect of status on dominant language and verbosity would be lower in the academic subpopulation than in the nonacademic cluster.</i>	2.98	4.02	4.52
<i>Users of the website increase their use of dominant language over time spent commenting</i>	2.77	4.67	4.67
<i>*Female participation correlates with number of females in discussion.</i>	3.63	5.10	5.85
<i>Male contributors will be less likely to refer to the work of others in their own responses</i>	3.34	4.23	5.08
<i>High status contributors coordinate their linguistic style less than low status contributors.</i>	3.41	5.22	5.13
<i>Male contributors coordinate their linguistic style less than female contributors.</i>	3.97	4.78	5.09
<i>Gender and status interact, such that high status is a better predictor of language coordination for male contributors than female contributors.</i>	3.27	4.4	4.98
<i>High-status individuals are more likely to introduce (self-declared) novel ideas in scientific conversations than low-status individuals.</i>	3.12	4.57	4.77

Note. * Indicates crowdsourced hypotheses that were chosen to be analyzed in the pilot study

The pilot, in which 12 research teams tested the final set of hypotheses (a number from the project coordinators, and several generated by the crowd as indicated above) is reported in Supplement 3. We believe this crowdsourced generation, evaluation, and testing (CGET) approach may be broadly applicable, and especially useful in cases of “closed” data that for legal or ethical reasons cannot be distributed outside a small team of investigators. Even under such constraints, a data descriptor with a variables list and descriptive statistics for the sample can be publicly posted and research ideas solicited from the community.

Appendix S2-1: Hypothesis generation survey

Hypothesis Generation Survey

Q1 You may have your own original hypothesis related to gender and status in scientific conversations. If you would like to nominate your hypothesis for testing using a crowdsourced data analysis approach, please state it here

Q2 Please list your arguments for why this hypothesis should be tested using a crowdsourced data analysis approach

Q3 Please list any scientific references relevant to this hypothesis

Q5 If you would like to be credited for your hypothesis, you can put your name here. If you prefer to be anonymous, you can leave this blank.

Q4 All of the nominated hypotheses will be sent to the entire group for peer evaluation, and the most favorably evaluated ideas will be selected for testing using the crowdsourcing analytics approach

Appendix S2-2: Hypothesis evaluation survey

Hypothesis Evaluation Survey

Q80 On the following pages, you will read seven additional hypotheses, some with multiple parts, that have been suggested to be included in our research project. We kindly ask that you evaluate each of them using the questions provided.

Q1 Additional Hypothesis 1: Male commenters will comment more frequently and at greater length than female commenters. This will be especially true for live conversations, and attenuated/non-existent for the non-conversation-format comments

Q29 Relevant literature: http://www.culturarsc.com/Genero/brescollv_who_takes_the_floor.pdf.

Q2 This hypothesis should be tested in this crowdsourcing data analysis project

- ☐ Strongly Disagree (1)
- ☐ Disagree (2)
- ☐ Somewhat Disagree (3)
- ☐ Neither Agree nor Disagree (4)
- ☐ Somewhat Agree (5)
- ☐ Agree (6)
- ☐ Strongly Agree (7)

Q3 Do you think this hypothesis is likely to be true?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q4 How scientifically interesting would it be if this hypothesis turned out to be true?

- ☐ Not at all scientifically interesting (1)
- ☐ Slightly scientifically interesting (2)
- ☐ Somewhat scientifically interesting (3)
- ☐ Very scientifically interesting (4)
- ☐ Extremely scientifically interesting (5)

Q57 How likely do you think it is that you could test this hypothesis with this dataset given your data analytic skills?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q41 Additional Hypothesis 2.1: The main effect of gender on dominant language and verbosity would be higher in the academic sub-population than in the non-academic cluster.

Q42 More information: I would like to use the power of the group to use Coarsened Exact Matching (CEM) to prepare the subgroups/clusters for testing, as this requires individual intervention ex ante. This reduces the error of one person deciding the criteria and matches.. Relevant Literature: <http://gking.harvard.edu/cem>; <http://gking.harvard.edu/files/abs/cemStata-abs.shtml>

Q43 This hypothesis should be tested in this crowdsourcing data analysis project

- ☐ Strongly Disagree (1)
- ☐ Disagree (2)
- ☐ Somewhat Disagree (3)
- ☐ Neither Agree nor Disagree (4)
- ☐ Somewhat Agree (5)
- ☐ Agree (6)
- ☐ Strongly Agree (7)

Q45 Do you think this hypothesis is likely to be true?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q46 How scientifically interesting would it be if this hypothesis turned out to be true?

- ☐ Not at all scientifically interesting (1)
- ☐ Slightly scientifically interesting (2)
- ☐ Somewhat scientifically interesting (3)
- ☐ Very scientifically interesting (4)
- ☐ Extremely scientifically interesting (5)

Q58 How likely do you think it is that you could test this hypothesis with this dataset given your data analytic skills?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q5 Additional Hypothesis 2.2: The main effect of status on dominant language and verbosity would be lower in the academic sub-population than in the non-academic cluster.

Q31 More information: I would like to use the power of the group to use Coarsened Exact Matching (CEM) to prepare the subgroups/clusters for testing, as this requires individual intervention ex ante. This reduces the error of one person deciding the criteria and matches.. Relevant Literature: <http://gking.harvard.edu/cem>; <http://gking.harvard.edu/files/abs/cemStata-abs.shtml>

Q6 This hypothesis should be tested in this crowdsourcing data analysis project

- ☐ Strongly Disagree (1)
- ☐ Disagree (2)
- ☐ Somewhat Disagree (3)
- ☐ Neither Agree nor Disagree (4)
- ☐ Somewhat Agree (5)
- ☐ Agree (6)
- ☐ Strongly Agree (7)

Q44 This hypothesis should be tested in this crowdsourcing data analysis project

- ☐ Strongly Disagree (1)
- ☐ Disagree (2)
- ☐ Somewhat Disagree (3)
- ☐ Neither Agree nor Disagree (4)
- ☐ Somewhat Agree (5)
- ☐ Agree (6)
- ☐ Strongly Agree (7)

Q7 Do you think this hypothesis is likely to be true?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q8 How scientifically interesting would it be if this hypothesis turned out to be true?

- ☐ Not at all scientifically interesting (1)
- ☐ Slightly scientifically interesting (2)
- ☐ Somewhat scientifically interesting (3)
- ☐ Very scientifically interesting (4)
- ☐ Extremely scientifically interesting (5)

Q59 How likely do you think it is that you could test this hypothesis with this dataset given your data analytic skills?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q9 Additional Hypothesis 3: Users of the website increase their use of dominant language over time spent commenting on Edge.org

Q30 More information: Having a variable as age of a contributor we can order their comments and find the point the use of dominant language reach a tipping point. Additionally, we can compare this to the years since first usage of the website to find out the number of years of usage until the use of dominant language.

Q10 This hypothesis should be tested in this crowdsourcing data analysis project

- ☐ Strongly Disagree (1)
- ☐ Disagree (2)
- ☐ Somewhat Disagree (3)
- ☐ Neither Agree nor Disagree (4)
- ☐ Somewhat Agree (5)
- ☐ Agree (6)
- ☐ Strongly Agree (7)

Q11 Do you think this hypothesis is likely to be true?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q12 How scientifically interesting would it be if this hypothesis turned out to be true?

- ☐ Not at all scientifically interesting (1)
- ☐ Slightly scientifically interesting (2)
- ☐ Somewhat scientifically interesting (3)
- ☐ Very scientifically interesting (4)
- ☐ Extremely scientifically interesting (5)

Q60 How likely do you think it is that you could test this hypothesis with this dataset given your data analytic skills?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q13 Additional Hypothesis 4: Female participation correlates with number of females in discussion.

Q32 More information: We believe this hypothesis should be considered because it has intrinsic social importance yet admits a variety of interpretations, suiting it well to a crowdsourced approach. On the first claim, the hypothesis is important as it might (if true) help conference or debate organizers to form panels, or educators to set up groups for group work. It could also serve as an argument to support quotas for women in boards etc. On the second claim, the measure of 'female participation' is not unique. For example, one might reasonably consider 'number of words contributed by women', 'number of distinct comments contributed by women', or 'number of distinct topics/ideas contributed by women.' The same goes for 'number of females in the discussion': perhaps it scales linearly, or perhaps it is having at least one other woman present that matters. If we crowdsource the analysis, different interpretations will surface naturally and hence strengthen the evidence for or against the hypothesis. Relevant Literature: <https://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=8675859&fulltextType=RA&fileId=S0003055412000329>

Q14 This hypothesis should be tested in this crowdsourcing data analysis project

- ☐ Strongly Disagree (1)
- ☐ Disagree (2)
- ☐ Somewhat Disagree (3)
- ☐ Neither Agree nor Disagree (4)
- ☐ Somewhat Agree (5)
- ☐ Agree (6)
- ☐ Strongly Agree (7)

Q15 Do you think this hypothesis is likely to be true?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q16 How scientifically interesting would it be if this hypothesis turned out to be true?

- ☐ Not at all scientifically interesting (1)
- ☐ Slightly scientifically interesting (2)
- ☐ Somewhat scientifically interesting (3)
- ☐ Very scientifically interesting (4)
- ☐ Extremely scientifically interesting (5)

Q61 How likely do you think it is that you could test this hypothesis with this dataset given your data analytic skills?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q17 Additional Hypothesis 5: Male contributors will be less likely to refer to the work of others in their own responses

Q33 More information: One way to think about "dominance" beyond language use is one's willingness to include and acknowledge the contributions of others.

Q18 This hypothesis should be tested in this crowdsourcing data analysis project

- ☐ Strongly Disagree (1)
- ☐ Disagree (2)
- ☐ Somewhat Disagree (3)
- ☐ Neither Agree nor Disagree (4)
- ☐ Somewhat Agree (5)
- ☐ Agree (6)
- ☐ Strongly Agree (7)

Q19 Do you think this hypothesis is likely to be true?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q20 How scientifically interesting would it be if this hypothesis turned out to be true?

- ☐ Not at all scientifically interesting (1)
- ☐ Slightly scientifically interesting (2)
- ☐ Somewhat scientifically interesting (3)
- ☐ Very scientifically interesting (4)
- ☐ Extremely scientifically interesting (5)

Q62 How likely do you think it is that you could test this hypothesis with this dataset given your data analytic skills?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q21 Additional Hypothesis 6.1: High status contributors coordinate their linguistic style less than low status contributors.

Q35 More information: Language coordination is the phenomenon in which people tend to unconsciously mimic others' linguistic style (e.g., use of first-person pronouns or prepositions) and has been shown to reveal power differentials in diverse contexts. For instance, in oral arguments for Supreme Court cases, lawyers mimic the linguistic style of Justices more than vice versa (Danescu-Niculescu-Mizil et al., 2012). Moreover, female lawyers were more likely than male lawyers to coordinate their style to the Justices, and Justices were less likely coordinate to female lawyers than to male lawyers. Analyzing language coordination can therefore help characterize power differentials and gender differences in intellectual discussions. Testing hypothesis about it could help inform the focal hypotheses that have been already proposed and vice versa. Measuring language coordination is tractable given the text data we have; for instance, see the analysis algorithms proposed by Danescu-Niculescu-Mizil et al., 2012. However, the idea is also broad enough that the different analysis teams will likely generate diverse, rich analysis strategies. Relevant Literature: Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. Proceedings of the 21st International Conference on the World Wide Web, pp. 699-708. Accessed December 15th, 2014 at <http://arxiv.org/pdf/1112.3670.pdf>.

Q22 This hypothesis should be tested in this crowdsourcing data analysis project

- ☐ Strongly Disagree (1)
- ☐ Disagree (2)
- ☐ Somewhat Disagree (3)
- ☐ Neither Agree nor Disagree (4)
- ☐ Somewhat Agree (5)
- ☐ Agree (6)
- ☐ Strongly Agree (7)

Q23 Do you think this hypothesis is likely to be true?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q24 How scientifically interesting would it be if this hypothesis turned out to be true?

- ☐ Not at all scientifically interesting (1)
- ☐ Slightly scientifically interesting (2)
- ☐ Somewhat scientifically interesting (3)
- ☐ Very scientifically interesting (4)
- ☐ Extremely scientifically interesting (5)

Q63 How likely do you think it is that you could test this hypothesis with this dataset given your data analytic skills?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q47 Additional Hypothesis 6.2: Male contributors coordinate their linguistic style less than female contributors.

Q48 More information: Language coordination is the phenomenon in which people tend to unconsciously mimic others' linguistic style (e.g., use of first-person pronouns or prepositions) and has been shown to reveal power differentials in diverse contexts. For instance, in oral arguments for Supreme Court cases, lawyers mimic the linguistic style of Justices more than vice versa (Danescu-Niculescu-Mizil et al., 2012). Moreover, female lawyers were more likely than male lawyers to coordinate their style to the Justices, and Justices were less likely coordinate to female lawyers than to male lawyers. Analyzing language coordination can therefore help characterize power differentials and gender differences in intellectual discussions. Testing hypothesis about it could help inform the focal hypotheses that have been already proposed and vice versa. Measuring language coordination is tractable given the text data we have; for instance, see the analysis algorithms proposed by Danescu-Niculescu-Mizil et al., 2012. However, the idea is also broad enough that the different analysis teams will likely generate diverse, rich analysis strategies. Relevant Literature: Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. Proceedings of the 21st International Conference on the World Wide Web, pp. 699-708. Accessed December 15th, 2014 at <http://arxiv.org/pdf/1112.3670.pdf>.

Q49 This hypothesis should be tested in this crowdsourcing data analysis project

- ☐ Strongly Disagree (1)
- ☐ Disagree (2)
- ☐ Somewhat Disagree (3)
- ☐ Neither Agree nor Disagree (4)
- ☐ Somewhat Agree (5)
- ☐ Agree (6)
- ☐ Strongly Agree (7)

Q50 Do you think this hypothesis is likely to be true?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q51 How scientifically interesting would it be if this hypothesis turned out to be true?

- ☐ Not at all scientifically interesting (1)
- ☐ Slightly scientifically interesting (2)
- ☐ Somewhat scientifically interesting (3)
- ☐ Very scientifically interesting (4)
- ☐ Extremely scientifically interesting (5)

Q66 How likely do you think it is that you could test this hypothesis with this dataset given your data analytic skills?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q74 Additional Hypothesis 6.2: Gender and status interact, such that high status is a better predictor of language coordination for male contributors than female contributors.

Q75 More information: Language coordination is the phenomenon in which people tend to unconsciously mimic others' linguistic style (e.g., use of first-person pronouns or prepositions) and has been shown to reveal power differentials in diverse contexts. For instance, in oral arguments for Supreme Court cases, lawyers mimic the linguistic style of Justices more than vice versa (Danescu-Niculescu-Mizil et al., 2012). Moreover, female lawyers were more likely than male lawyers to coordinate their style to the Justices, and Justices were less likely coordinate to female lawyers than to male lawyers. Analyzing language coordination can therefore help characterize power differentials and gender differences in intellectual discussions. Testing hypothesis about it

could help inform the focal hypotheses that have been already proposed and vice versa. Measuring language coordination is tractable given the text data we have; for instance, see the analysis algorithms proposed by Danescu-Niculescu-Mizil et al., 2012. However, the idea is also broad enough that the different analysis teams will likely generate diverse, rich analysis strategies. Relevant Literature: Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. Proceedings of the 21st International Conference on the World Wide Web, pp. 699-708. Accessed December 15th, 2014 at <http://arxiv.org/pdf/1112.3670.pdf>.

Q76 This hypothesis should be tested in this crowdsourcing data analysis project

- ☐ Strongly Disagree (1)
- ☐ Disagree (2)
- ☐ Somewhat Disagree (3)
- ☐ Neither Agree nor Disagree (4)
- ☐ Somewhat Agree (5)
- ☐ Agree (6)
- ☐ Strongly Agree (7)

Q77 Do you think this hypothesis is likely to be true?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q78 How scientifically interesting would it be if this hypothesis turned out to be true?

- ☐ Not at all scientifically interesting (1)
- ☐ Slightly scientifically interesting (2)
- ☐ Somewhat scientifically interesting (3)
- ☐ Very scientifically interesting (4)
- ☐ Extremely scientifically interesting (5)

Q79 How likely do you think it is that you could test this hypothesis with this dataset given your data analytic skills?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q25 Additional Hypothesis 7: High-status individuals are more likely to introduce (self-declared) novel ideas in scientific conversations than low-status individuals.

Q34 More information: Since it orients the research agendas of many individual scientists in new directions, steering the scientific debate towards novel questions and concepts has major ramifications. As anecdotal evidence on keynote speeches at scientific conferences indicate, such effects also apply to relatively minor suggestions and not only to major paradigm shifts (Kuhn, 1962, 2012). While prior research has looked into agenda-setting for public policy (e.g. Dearing & Rogers, 1996), it is not well-known how this process works in the scientific discourse. More specifically, due to lack of rich, structured micro-data about scientific discourse, it is unclear to what extent scientific agenda-setting conforms to a 'Matthew process' (Merton, 1968) in which the select few individuals who enjoy the highest status levels are the ones who are in the driver's seat, or whether it is a more egalitarian process. Relevant Literature: Dearing, J. W., & Rogers, E. M. (1996). Agenda-setting (Vol. 6). Sage Publications. Kuhn, T. S. (2012). The structure of scientific revolutions. University of Chicago press. Merton, R. K. (1968). The Matthew effect in science. Science, 159(3810), 56-63.

Q26 This hypothesis should be tested in this crowdsourcing data analysis project

- ☐ Strongly Disagree (1)
- ☐ Disagree (2)
- ☐ Somewhat Disagree (3)
- ☐ Neither Agree nor Disagree (4)
- ☐ Somewhat Agree (5)
- ☐ Agree (6)
- ☐ Strongly Agree (7)

Q27 Do you think this hypothesis is likely to be true?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Q28 How scientifically interesting would it be if this hypothesis turned out to be true?

- ☐ Not at all scientifically interesting (1)
- ☐ Slightly scientifically interesting (2)
- ☐ Somewhat scientifically interesting (3)
- ☐ Very scientifically interesting (4)
- ☐ Extremely scientifically interesting (5)

Q64 How likely do you think it is that you could test this hypothesis with this dataset given your data analytic skills?

- ☐ Very Unlikely (1)
- ☐ Unlikely (2)
- ☐ Somewhat Unlikely (3)
- ☐ Undecided (4)
- ☐ Somewhat Likely (5)
- ☐ Likely (6)
- ☐ Very Likely (7)

Supplement 3: Pilot study

In the pilot, 18 research teams who expressed interest based on advertisements on social media and email lists (e.g., Twitter, Facebook, Academy of Management list serves for various interest groups) completed a pre-survey (see Appendix S3-1 and <https://osf.io/wd67q/>), then downloaded the Edge dataset to test eleven hypotheses. Twelve teams completed all of their analyses. Table S3-1 summarizes the number of analysis teams that obtained significant support for each hypothesis, directional but non-significant support, results directionally opposite to the hypothesis, and significant results directly contrary to the original prediction. Also included in Table S3-1 are counts of the number of teams that conducted analyses but did not clearly report their results, as well as teams that conducted multiple analyses without clearly identifiable results. Only teams that clearly reported a single result for a given hypothesis are included in the significance and direction counts.

As shown in the table, none of the eleven hypotheses tested in the pilot enjoyed quantitative support at the $p < .05$ level based on the analyses of more than 5 out of 12 research teams. Indeed, in several cases different research teams obtained statistically significant effects in opposing directions despite testing the same research question with the same data.

In our primary study, we focused on just two hypotheses that exhibited highly dispersed results across different analysts, collected more details to allow the coordination team to error check and reproduce the analyses, and introduced the DataExplained platform to capture researchers' rationales for each step undertaken during the data analysis process. This, we hoped, would allow us to better understand why empirical results can be so different across different investigators using the same set of observations to test the same idea. We thus sought to replicate the key results of the pilot while collecting more documentation and process measures.

Table S3-1: *Direction and significance levels for each hypothesis across different analysis teams in the pilot study*

Hypothesis	Significant in predicted (+) direction	Not significant in predicted (+) direction	Not significant in opposite (-) direction	Significant in opposite (-) direction	Number of analyses without results	Number of analyses with multiple, not clearly identifiable results
<i>Higher status participants are more verbose than are lower status participants.</i>	12.5% (n=1)	37.5% (n=3)	12.5% (n=1)	37.5% (n=3)	(n=3)	(n=1)
<i>Higher status participants use more dominant language than do lower status participants.</i>	25% (n=2)	25% (n=2)	50% (n=4)	0% (n=0)	(n=2)	(n=2)
<i>Male participants are more verbose than female participants.</i>	22.2% (n=2)	33.3% (n=3)	44.4% (n=4)	0% (n=0)	(n=1)	(n=2)
<i>Male participants use more dominant language than do female participants.</i>	0% (n=0)	66.6% (n=6)	33.3% (n=3)	0% (n=0)	(n=2)	(n=1)
<i>Gender and status interact, such that high status is a better predictor of verbosity for male scientists than for female scientists.</i>	0% (n=0)	71.4% (n=5)	28.6% (n=2)	0% (n=0)	(n=1)	(n=4)
<i>Gender and status interact, such that high status is a better predictor of dominant language among male scientists than for female scientists.</i>	0% (n=0)	57.1% (n=4)	42.9% (n=3)	0% (n=0)	(n=3)	(n=2)

<i>Very low and very high status participants are the least likely to be verbose.</i>	25% (n=2)	25% (n=2)	50% (n=4)	0% (n=0)	(n=3)	(n=1)
<i>Very low and very high status participants are the least likely to use dominant language.</i>	0% (n=0)	62.5% (n=5)	12.5% (n=1)	25% (n=2)	(n=4)	(n=0)
<i>Female participation correlates with the number of females in the discussion.</i>	50% (n=5)	30% (n=3)	10% (n=1)	10% (n=1)	(n=2)	(n=0)
<i>The effect of gender on verbosity will be the strongest in live conversations and attenuated for the asynchronous format conversations.</i>	14.3% (n=1)	14.3% (n=1)	71.4% (n=5)	0% (n=0)	(n=2)	(n=3)
<i>The effect of gender on use of dominant language will be the strongest in live conversations and attenuated for the asynchronous format conversations.</i>	0% (n=0)	66.7% (n=4)	0% (n=0)	33.3% (n=2)	(n=3)	(n=3)

Note. Percentage terms were calculated based on the number of analyses for which direction and statistical significance levels were known

Appendix S3-1: Pre-survey from pilot study

Q27 What is your full name?

Q1 My Gender is:

- ☐ Male (1)
- ☐ Female (2)
- ☐ Other (3) _____

Q29 What year were you born?

- ☐ 1920 (1)
- ☐ 1921 (2)
- ☐ 1922 (3)
- ☐ 1923 (4)
- ☐ 1924 (5)
- ☐ 1925 (6)
- ☐ 1926 (7)
- ☐ 1927 (8)
- ☐ 1928 (9)
- ☐ 1929 (10)
- ☐ 1930 (11)
- ☐ 1931 (12)
- ☐ 1932 (13)
- ☐ 1933 (14)
- ☐ 1934 (15)
- ☐ 1935 (16)
- ☐ 1936 (17)
- ☐ 1937 (18)
- ☐ 1938 (19)
- ☐ 1939 (20)
- ☐ 1940 (21)
- ☐ 1941 (22)
- ☐ 1942 (23)
- ☐ 1943 (24)
- ☐ 1944 (25)
- ☐ 1945 (26)
- ☐ 1946 (27)
- ☐ 1947 (28)
- ☐ 1948 (29)
- ☐ 1949 (30)
- ☐ 1950 (31)
- ☐ 1951 (32)
- ☐ 1952 (33)
- ☐ 1953 (34)
- ☐ 1954 (35)
- ☐ 1955 (36)

- ☐ 1956 (37)
- ☐ 1957 (38)
- ☐ 1958 (39)
- ☐ 1959 (40)
- ☐ 1960 (41)
- ☐ 1961 (42)
- ☐ 1962 (43)
- ☐ 1963 (44)
- ☐ 1964 (45)
- ☐ 1965 (46)
- ☐ 1966 (47)
- ☐ 1967 (48)
- ☐ 1968 (49)
- ☐ 1969 (50)
- ☐ 1970 (51)
- ☐ 1971 (52)
- ☐ 1972 (53)
- ☐ 1973 (54)
- ☐ 1974 (55)
- ☐ 1975 (56)
- ☐ 1976 (57)
- ☐ 1977 (58)
- ☐ 1978 (59)
- ☐ 1979 (60)
- ☐ 1980 (61)
- ☐ 1981 (62)
- ☐ 1982 (63)
- ☐ 1983 (64)
- ☐ 1984 (65)
- ☐ 1985 (66)
- ☐ 1986 (67)
- ☐ 1987 (68)
- ☐ 1988 (69)
- ☐ 1989 (70)
- ☐ 1990 (71)
- ☐ 1991 (72)
- ☐ 1992 (73)
- ☐ 1993 (74)
- ☐ 1994 (75)
- ☐ 1995 (76)
- ☐ 1996 (77)
- ☐ 1997 (78)
- ☐ 1998 (79)
- ☐ 1999 (80)
- ☐ 2000 (81)

Q18 What is your highest degree?

- ☐ PhD (1)
- ☐ Master's (2)
- ☐ Bachelor's (3)

Q15 If you have a PhD, what field is it in?

Q20 What title best describes your current position?

- ☐ Full Professor (1)
- ☐ Associate Professor (2)
- ☐ Assistant Professor (3)
- ☐ Post-Doc (4)
- ☐ Doctoral Student (5)
- ☐ Other position at a University (6) _____
- ☐ Outside of Academia (7) _____

Q6 In which country were you born?

- ☐ Afghanistan (1)
- ☐ Albania (2)
- ☐ Algeria (3)
- ☐ Andorra (4)
- ☐ Angola (5)
- ☐ Antigua and Barbuda (6)
- ☐ Argentina (7)
- ☐ Armenia (8)
- ☐ Australia (9)
- ☐ Austria (10)
- ☐ Azerbaijan (11)
- ☐ Bahamas (12)
- ☐ Bahrain (13)
- ☐ Bangladesh (14)
- ☐ Barbados (15)
- ☐ Belarus (16)
- ☐ Belgium (17)
- ☐ Belize (18)
- ☐ Benin (19)
- ☐ Bhutan (20)
- ☐ Bolivia (21)
- ☐ Bosnia and Herzegovina (22)
- ☐ Botswana (23)
- ☐ Brazil (24)
- ☐ Brunei Darussalam (25)
- ☐ Bulgaria (26)
- ☐ Burkina Faso (27)
- ☐ Burundi (28)
- ☐ Cambodia (29)

- Cameroon (30)
- Canada (31)
- Cape Verde (32)
- Central African Republic (33)
- Chad (34)
- Chile (35)
- China (36)
- Colombia (37)
- Comoros (38)
- Congo, Republic of the... (39)
- Costa Rica (40)
- Côte d'Ivoire (41)
- Croatia (42)
- Cuba (43)
- Cyprus (44)
- Czech Republic (45)
- Democratic People's Republic of Korea (46)
- Democratic Republic of the Congo (47)
- Denmark (48)
- Djibouti (49)
- Dominica (50)
- Dominican Republic (51)
- Ecuador (52)
- Egypt (53)
- El Salvador (54)
- Equatorial Guinea (55)
- Eritrea (56)
- Estonia (57)
- Ethiopia (58)
- Fiji (59)
- Finland (60)
- France (61)
- Gabon (62)
- Gambia (63)
- Georgia (64)
- Germany (65)
- Ghana (66)
- Greece (67)
- Grenada (68)
- Guatemala (69)
- Guinea (70)
- Guinea-Bissau (71)
- Guyana (72)
- Haiti (73)
- Honduras (74)
- Hong Kong (S.A.R.) (75)

- Hungary (76)
- Iceland (77)
- India (78)
- Indonesia (79)
- Iran, Islamic Republic of... (80)
- Iraq (81)
- Ireland (82)
- Israel (83)
- Italy (84)
- Jamaica (85)
- Japan (86)
- Jordan (87)
- Kazakhstan (88)
- Kenya (89)
- Kiribati (90)
- Kuwait (91)
- Kyrgyzstan (92)
- Lao People's Democratic Republic (93)
- Latvia (94)
- Lebanon (95)
- Lesotho (96)
- Liberia (97)
- Libyan Arab Jamahiriya (98)
- Liechtenstein (99)
- Lithuania (100)
- Luxembourg (101)
- Madagascar (102)
- Malawi (103)
- Malaysia (104)
- Maldives (105)
- Mali (106)
- Malta (107)
- Marshall Islands (108)
- Mauritania (109)
- Mauritius (110)
- Mexico (111)
- Micronesia, Federated States of... (112)
- Monaco (113)
- Mongolia (114)
- Montenegro (115)
- Morocco (116)
- Mozambique (117)
- Myanmar (118)
- Namibia (119)
- Nauru (120)
- Nepal (121)

- Netherlands (122)
- New Zealand (123)
- Nicaragua (124)
- Niger (125)
- Nigeria (126)
- North Korea (127)
- Norway (128)
- Oman (129)
- Pakistan (130)
- Palau (131)
- Panama (132)
- Papua New Guinea (133)
- Paraguay (134)
- Peru (135)
- Philippines (136)
- Poland (137)
- Portugal (138)
- Qatar (139)
- Republic of Korea (140)
- Republic of Moldova (141)
- Romania (142)
- Russian Federation (143)
- Rwanda (144)
- Saint Kitts and Nevis (145)
- Saint Lucia (146)
- Saint Vincent and the Grenadines (147)
- Samoa (148)
- San Marino (149)
- Sao Tome and Principe (150)
- Saudi Arabia (151)
- Senegal (152)
- Serbia (153)
- Seychelles (154)
- Sierra Leone (155)
- Singapore (156)
- Slovakia (157)
- Slovenia (158)
- Solomon Islands (159)
- Somalia (160)
- South Africa (161)
- South Korea (162)
- Spain (163)
- Sri Lanka (164)
- Sudan (165)
- Suriname (166)
- Swaziland (167)

- ☐ Sweden (168)
- ☐ Switzerland (169)
- ☐ Syrian Arab Republic (170)
- ☐ Tajikistan (171)
- ☐ Thailand (172)
- ☐ The former Yugoslav Republic of Macedonia (173)
- ☐ Timor-Leste (174)
- ☐ Togo (175)
- ☐ Tonga (176)
- ☐ Trinidad and Tobago (177)
- ☐ Tunisia (178)
- ☐ Turkey (179)
- ☐ Turkmenistan (180)
- ☐ Tuvalu (181)
- ☐ Uganda (182)
- ☐ Ukraine (183)
- ☐ United Arab Emirates (184)
- ☐ United Kingdom of Great Britain and Northern Ireland (185)
- ☐ United Republic of Tanzania (186)
- ☐ United States of America (187)
- ☐ Uruguay (188)
- ☐ Uzbekistan (189)
- ☐ Vanuatu (190)
- ☐ Venezuela, Bolivarian Republic of... (191)
- ☐ Viet Nam (192)
- ☐ Yemen (193)
- ☐ Zambia (580)
- ☐ Zimbabwe (1357)

Q2 In which country do you reside?

- ☐ Please select below... (1)
- ☐ Afghanistan (2)
- ☐ Albania (3)
- ☐ Algeria (4)
- ☐ Andorra (5)
- ☐ Angola (6)
- ☐ Antigua and Barbuda (7)
- ☐ Argentina (8)
- ☐ Armenia (9)
- ☐ Australia (10)
- ☐ Austria (11)
- ☐ Azerbaijan (12)
- ☐ Bahamas (13)
- ☐ Bahrain (14)
- ☐ Bangladesh (15)
- ☐ Barbados (16)

- Belarus (17)
- Belgium (18)
- Belize (19)
- Benin (20)
- Bhutan (21)
- Bolivia (22)
- Bosnia and Herzegovina (23)
- Botswana (24)
- Brazil (25)
- Brunei (26)
- Bulgaria (27)
- Burkina Faso (28)
- Burma/Myanmar (29)
- Burundi (30)
- Cambodia (31)
- Cameroon (32)
- Canada (33)
- Cape Verde (34)
- Central African Republic (35)
- Chad (36)
- Chile (37)
- China (38)
- Colombia (39)
- Comoros (40)
- Congo (41)
- Congo, Democratic Republic of (42)
- Costa Rica (43)
- Cote d'Ivoire/Ivory Coast (44)
- Croatia (45)
- Cuba (46)
- Cyprus (47)
- Czech Republic (48)
- Denmark (49)
- Djibouti (50)
- Dominica (51)
- Dominican Republic (52)
- East Timor (53)
- Ecuador (54)
- Egypt (55)
- El Salvador (56)
- Equatorial Guinea (57)
- Eritrea (58)
- Estonia (59)
- Ethiopia Fiji (60)
- Finland (61)
- France (62)

- Gabon (63)
- Gambia (64)
- Georgia (65)
- Germany (66)
- Ghana (67)
- Greece (68)
- Grenada (69)
- Guatemala (70)
- Guinea (71)
- Guinea-Bissau (Bissau) (AF) (72)
- Guyana (73)
- Haiti (74)
- Honduras (75)
- Hungary (76)
- Iceland (77)
- India (78)
- Indonesia (79)
- Iran (80)
- Iraq (81)
- Ireland (82)
- Israel (83)
- Italy (84)
- Jamaica (85)
- Japan (86)
- Jordan (87)
- Kazakhstan (88)
- Kenya (89)
- Kiribati (90)
- Korea, North (91)
- Korea, South (92)
- Kuwait (93)
- Kyrgyzstan (94)
- Laos (95)
- Latvia (96)
- Lebanon (97)
- Lesotho (98)
- Liberia (99)
- Libya (100)
- Liechtenstein (101)
- Lithuania (102)
- Luxembourg (103)
- Macedonia (104)
- Madagascar (105)
- Malawi (106)
- Malaysia (107)
- Maldives (108)

- Mali (109)
- Malta (110)
- Marshall Islands (111)
- Mauritania (112)
- Mauritius (113)
- Mexico (114)
- Micronesia (115)
- Moldova (116)
- Monaco (117)
- Mongolia (118)
- Montenegro (119)
- Morocco (120)
- Mozambique (121)
- Namibia (122)
- Nauru (123)
- Nepal (124)
- Netherlands (125)
- New Zealand (126)
- Nicaragua (127)
- Niger (128)
- Nigeria (129)
- Norway (130)
- Oman (131)
- Pakistan (132)
- Palau (133)
- Panama (134)
- Papua New Guinea (135)
- Paraguay (136)
- Peru (137)
- Philippines (138)
- Poland (139)
- Portugal (140)
- Qatar (141)
- Romania (142)
- Russian Federation (143)
- Rwanda (144)
- Saint Kitts and Nevis (145)
- Saint Lucia (146)
- Saint Vincent and the Grenadines (147)
- Samoa (148)
- San Marino (149)
- Sao Tome and Principe (150)
- Saudi Arabia (151)
- Senegal (152)
- Serbia (153)
- Seychelles (154)

- Sierra Leone (155)
- Singapore (156)
- Slovakia (157)
- Slovenia (158)
- Solomon Islands (159)
- Somalia (160)
- South Africa (161)
- Spain (162)
- Sri Lanka (163)
- Sudan (164)
- Suriname (165)
- Swaziland (166)
- Sweden (167)
- Switzerland (168)
- Syria (169)
- Taiwan (170)
- Tajikistan (171)
- Tanzania (172)
- Thailand (173)
- Togo (174)
- Tonga (175)
- Trinidad and Tobago (176)
- Tunisia (177)
- Turkey (178)
- Turkmenistan (179)
- Tuvalu (180)
- Uganda (181)
- Ukraine (182)
- United Arab Emirates (183)
- United Kingdom (184)
- United States (185)
- Uruguay (186)
- Uzbekistan (187)
- Vanuatu (188)
- Vatican City (189)
- Venezuela (190)
- Vietnam (191)
- Yemen (192)
- Zambia (193)
- Zimbabwe (194)
- Other (195)

Q3 Please rate your political ideology on the following scale:

- ☐ Strongly Left-Wing (1)
- ☐ Moderately Left-Wing (2)
- ☐ Slightly Left Wing (3)
- ☐ Moderate (4)
- ☐ Slightly Right Wing (5)
- ☐ Moderately Right-Wing (6)
- ☐ Strongly Right-Wing (7)

Q26 Below you will find a set of skills and behaviors that you or your team, if you are collaborating with people to perform the analyses, will engage in while conducting the analyses. Please indicate how confident you are that you or your team is able to do the following (from 1=*Cannot do at all*, to 100=*Highly certain can do*):

- _____ Operationalize key variables based on theoretically defensible rationales (1)
- _____ Handle a large data set (2)
- _____ Use appropriate analytic techniques to test the proposed hypotheses (3)
- _____ Provide a clear description of the analysis strategy and rationale (4)

Q22 Have you taught an undergraduate level statistics course? If so, how many total times (estimate is fine)?

- ☐ 0 (1)
- ☐ 1-2 (2)
- ☐ 3-5 (3)
- ☐ more than 5 (4)

Q24 Have you taught an undergraduate level course on analyzing text? If so, how many total times (estimate is fine)?

- ☐ 0 (1)
- ☐ 1-2 (2)
- ☐ 3-5 (3)
- ☐ more than 5 (4)

Q24 Have you taught a graduate level statistics course? If so, how many total times (estimate is fine)?

- ☐ 0 (1)
- ☐ 1-2 (2)
- ☐ 3-5 (3)
- ☐ more than 5 (4)

Q23 Have you taught a graduate level course on analyzing text? If so, how many total times (estimate is fine)?

- ☐ 0 (1)
- ☐ 1-2 (2)
- ☐ 3-5 (3)
- ☐ more than 5 (4)

Q5 Approximately how many Edge conversations have you read before?

- ☐ 0 (1)
- ☐ 1-10 (2)
- ☐ 11-20 (3)
- ☐ 20-50 (4)
- ☐ more than 50 (5)

Q6 Have you published a peer-reviewed, scientific paper on text analysis, using a statistical analysis of text? If not, please put 0, and if so, please put how many (estimate is fine)?

- ☐ 0 (1)
- ☐ 1-2 (2)
- ☐ 3-5 (3)
- ☐ 6-8 (4)
- ☐ more than 8 (5)

Q8 Have you published a peer-reviewed, scientific paper that is on the topic of gender? If not, please put 0, and if so, please put how many (estimate is fine)? If you have published articles on both both gender and status, please include it in the gender and the status publication counts.

- ☐ 0 (1)
- ☐ 1-2 (2)
- ☐ 3-5 (3)
- ☐ 6-8 (4)
- ☐ more than 8 (5)

Q10 Have you published a peer-reviewed, scientific article on the topic of social status? If not, please put 0, and if so, please put how many (estimate is fine)? If you have published articles on both both gender and status, please include it in the gender and the status publication counts.

- ☐ 0 (1)
- ☐ 1-2 (2)
- ☐ 3-5 (3)
- ☐ 6-8 (4)
- ☐ more than 8 (5)

Q26 Have you published a paper that is primarily a methodological/statistical contribution? If not, please put 0, and If so, how many in total (estimate is fine)?

- ☐ 0 (1)
- ☐ 1-2 (2)
- ☐ 3-5 (3)
- ☐ 6-8 (4)
- ☐ more than 8 (5)

Supplement 4: Online advertisements for primary project

The advertisement for Facebook and other social media was posted in the PsychMAP group, Psych Methods Discussion group, R users group, and on the labinthewild website.

TWITTER ADVERTISEMENT

Become a coauthor by examining how analytic decisions affect research results re “gender, status, and science” <https://goo.gl/bnVVfS>

POSTING FOR FACEBOOK AND OTHER SOCIAL MEDIA SITES

Crowdsourcing Data Analysis 2, Phase 2: Explaining Variability in Analyses and Results

Interested in how analytic choices affect research results? Interested in the role of gender in scientific debates? Do you know how to analyse data using R? Join us as an analyst and co-author for the second phase of our project crowdsourcing the analysis of a dataset on gender, status, and science. If interested, please email Martin Schweinsberg (martin.schweinsberg@gmail.com) and Michael Feldman (feldman@ifi.uzh.ch). For a more detailed project description, click on this link: [<https://docs.google.com/document/d/1fXQBLdWydISskOKhoq8gl5unuwsv7VA3pkKY4IWFS6o/edit>]

FULL PROJECT DESCRIPTION IN ONLINE GOOGLE DOC

Crowdsourcing Data Analysis 2, Phase 2: Explaining Variability in Analyses and Results

Interested in how analytic choices affect research results? Interested in the role of gender in scientific debates? Do you know how to analyse data using R? Join us as an analyst and co-author for the second phase of our project crowdsourcing the analysis of a dataset on gender, status, and science.

We are employing the new approach of crowdsourcing data analysis, in which many independent analysts are recruited to test the same hypotheses on the same data set. Our first crowdsourcing data analysis initiative examined whether soccer referees give more red cards to dark skin toned than light-skin toned players (Silberzahn et al., under review; see project page on the Open Science Framework at <https://osf.io/gvm2z/>). The outcome was striking: although approximately two-thirds of teams obtained a significant effect in the expected direction, estimated effect sizes ranged from moderately large to practically nil.

In this second project “Crowdsourcing Data Analysis 2: Science, Gender, and Status” we are in the process of crowdsourcing the analysis of a dataset on intellectual conversations. You can become a co-author on the project by conducting analyses to test two hypotheses regarding the roles of the speaker’s status and gender in debates between scientists.

Critically, in this new phase of Crowdsourcing Data Analysis 2 we are trying to pinpoint exactly WHY analytic choices have such a profound effect on research results. We are recruiting scientists who analyse their data using R and are willing to use our new “Data Explained” platform to carefully track their analytic decisions in real time. Using Data Explained, we hope to identify the factors that play a role in data analysis variability. You can see a video tutorial for the platform here: <https://www.youtube.com/watch?v=UVNIFJaeNwI>

What is the Dataset?

- The data comes from Edge.org, a platform for intellectual discussion and debate
- The more than 700 contributors are chosen by Edge based on their creative work and include Daniel Kahneman, Marissa Meyer, Craig Venter, and many other academics as well as writers, entrepreneurs, business leaders, and more
- Each row is one comment in a conversation. There are approximately 7,600 rows and 3.8 million words. We also have information about the contributors.

What are the two hypotheses?

- Hypothesis 1: Female participation correlates with the number of females in the discussion
- Hypothesis 2: Higher status participants are more verbose than are lower status participants
- For those less familiar with text analysis, the following resources could be helpful: <https://discovertext.com/>, <http://www.liwc.net/>, <http://www.uclassify.com/>

Information for Collaborators

- Project participants must work alone for the analysis phase of the project so that we can track your decisions using the Data Explained platform. *During the analysis phase of the project, you should conduct your analysis independently, without collaborating or corresponding regarding your analysis with other project participants.* At a later stage in the project (see timeline below) you will be able to discuss your analyses with each other.
- Every person who completes and submits his or her analyses using the Data Explained platform, and subjects his or her final analysis report within the stated timeframe will be an author on the final paper (listed in alphabetical order after the coordination team and before the senior and last author).
- Each project contribution must include: (1) the code for the analysis and specification of analysis package required to execute the analysis, (2) a description of the rationale for the analysis strategy via Data Explained, (3) a complete written summary of the analysis strategy, and (4) a description of the result including specification of the effect estimate in effect size units (d , r , R^2 or odds ratio) and confidence interval.
- We are looking for colleagues with a wide range of expertise to participate in this crowdsourcing project, including researchers interested in text analysis, time, gender, status, and statistics. All participants must be able to conduct their analyses in R so they can use the Data Explained platform.
- **If you would like to join this study as an analyst and collaborator, please contact Martin Schweinsberg (martin.schweinsberg@gmail.com) and Michael Feldman (feldman@ifi.uzh.ch)**

Timeline:

- May 15 - June 30 2017: Analysis phase-- contributors analyze the dataset using the Data Explained platform
- July 1st 2017: Contributors submit their analytic approaches
- July 1st – August 1st 2017: Coordinators compile the results
- August 1st-August 31st 2017: Online discussion of the project results
- August 31st-November 31st 2017: Writeup of the final project report and comments from contributors via Google doc
- December 2017: Submission of the final report to a top academic journal for publication

Supplement 5: Pre-survey for analysts in primary study

Note: The pre-survey for the primary study and pilot study (Supplement 3) both included individual-differences measures of views on gender and political issues. Although the sample sizes of analysts are small, colleagues who wish to conduct exploratory analyses of the relationships between individual differences in beliefs and empirical results of the crowd analyses can do so with the publicly posted data (<https://osf.io/y9fq4/>).

Crowdsourcing Data Analysis 2: Explaining Variability in Data Analysis Decisions

Dear colleague,

Thank you for joining us as a collaborator and co-author on this crowdsourced project. Your work on the dataset and answers in this survey will help us better understand the reasons for variability in data analytic choices. Before embarking on the project, we would like to ask you a few questions about your background and experience.

The survey consists of 39 questions and will not take more than 15-20 minutes to answer.

Your responses, along with those from other project collaborators, will be used only for scholarly purposes and will be kept anonymous (i.e., will not be associated with your name).

Please note that authorship on the final project report is contingent on completing all stages of the project, including not only this presurvey but also the analysis of the dataset and tracking your decisions using the DataExplained process.

Q1: What is your name?

Q2: What is your username for DataExplained?

Q3: What is your highest degree?

- PhD
- Master's
- Bachelor's
- Other (Textfield)

Q4: What field is your PhD in? (Textfield) (*only displayed when PhD was selected in Q3*)

Q5: Please explain your professional background: e.g. Bachelor in Psychology, Master in Cognitive Psychology

Q6: In which country were you born? (Dropdown of countries)

Q7: In which country do you reside? (Dropdown of countries)

Q8: Please rate your political ideology on the following scale:

- Strongly Left-Wing
- Moderately Left-Wing
- Slightly Left-Wing
- Moderate
- Slightly Right-Wing
- Moderately Right-Wing
- Strongly Right-Wing

Q9: What are the keywords best describing the topics of your research?

Q10: What language/software/tools/do you prefer using in your works when doing data analysis? (e.g. R, STATA, Python, ...)

Q11: How many years of experience do you have in data analysis?

Q12: How regularly do you perform data analysis?

- Daily
- 2-3 times a week
- Once a week
- Once every two weeks
- Once a month
- Less than once a month

Q13: Please explain your background in data analysis in more detail: (e.g. what classes did you take, what projects have you analyzed etc.)

Q14: Below you will find a set of skills and behaviors that you likely engage in while conducting the analysis. Please indicate how confident (%) you are that you are able to do the following: (0 - Cannot do at all, 50 - Moderately can do, 100 - Highly certain can do)

- Operationalize key variables based on theoretically defensible rationales
- Handle a large data set
- Use appropriate analytic techniques to test the proposed hypotheses
- Provide a clear description of the analysis strategy and rationale

Q15: Below is a list of statistical methods. To what extent do you consider yourself to be skilled in each of them? (Not at all, To a low extend, To a medium extend, To a high extend, To a very high extend)

- Descriptive statistics (for example median or variance)
- Inferential statistics
- Hypothesis
- Ingression
- Estimation
- Correlation
- Regressions
- Forecasting
- Prediction
- Extrapolation
- Interpolation
- Time series
- Data mining

Q16: How do you rate your level of expertise in the field of data analysis?

1. Very poor
2. Amateur
3. Good
4. Very Good
5. Excellent

Q17: Have you taught an undergraduate level statistics course? If so, how many total times (estimate is fine)?

- 0
- 1-2
- 3-5
- more than 5

Q18: Have you taught an undergraduate level course on analyzing text? If so, how many total times (estimate is fine)?

- 0
- 1-2
- 3-5
- more than 5

Q19: Have you taught a graduate level statistics course? If so, how many total times (estimate is fine)?

- 0
- 1-2
- 3-5
- more than 5

Q20: Have you taught a graduate level course on analyzing text? If so, how many total times (estimate is fine)?

- 0
- 1-2
- 3-5
- more than 5

Q21: Approximately how many Edge conversations have you read before? Edge.org is an online website for intellectual discussions.

- 0
- 1-2
- 3-5
- more than 5

Q22: Have you published a peer-reviewed, scientific paper using text analysis? If not, please put 0, and if so, please put how many (estimate is fine):

- 0
- 1-2
- 3-5
- 6-8
- more than 8

Q23: Have you published a peer-reviewed, scientific paper that is on the topic of gender? If not, please put 0, and if so, please put how many (estimate is fine). If you have published articles on both gender AND status, please include it in the gender and the status publication counts.

- 0
- 1-2
- 3-5
- 6-8
- more than 8

Q24: Have you published a peer-reviewed, scientific paper that is on the topic of social status? If not, please put 0, and if so, please put how many (estimate is fine). If you have published articles on both gender AND status, please include it in the gender and the status publication counts.

- 0
- 1-2
- 3-5
- 6-8
- more than 8

Q25: Have you published a paper that is primarily a methodological/statistical contribution? If not, please put 0, and If so, how many in total (estimate is fine)?

- 0
- 1-2
- 3-5
- 6-8
- more than 8

Q26: To what extent does your research focus on social status? (7 point Likert scale with 1 = “Not at all” and 7 = “Extremely”)

Q27: Please briefly tell us about this research (Textfield)

Q28: In your personal opinion and experience, to what extent do you find a person’s status to play a role in their professional interactions in science? (7 point Likert scale with 1 = “Not at all” and 7 = “Extremely”)

Q29: Please briefly tell us about this: (Textfield) (*Only displayed if answer to previous question was 2 or more*)

Q30: To what extent does your research focus on gender issues? (7 point Likert scale with 1 = “Not at all” and 7 = “Extremely”)

Q31: In your personal opinion and experience, to what extent do you find a person’s gender to play a role in their scientific career? (7 point Likert scale with 1 = “Not at all” and 7 = “Extremely”)

Q32: Please briefly tell us about this: (Textfield) (*Only displayed if answer to previous question was 2 or more*)

Q33: What is your current opinion regarding hypothesis 1: A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion.

- Very Unlikely
- Unlikely
- Neither Likely nor Unlikely
- Likely
- Very Likely

Q34: Please explain why you think so: (Textfield)

Q35: What is your current opinion regarding hypothesis 2: Higher status participants are more verbose than are lower status participants.

- Very Unlikely
- Unlikely
- Neither Likely nor Unlikely
- Likely
- Very Likely

Q36: Please explain why you think so: (Textfield)

Q37: What is your gender?

- Female
- Male
- Other (Textfield)

Q38: What is your age? (Textfield)

Q39: What title best describes your current position?

- Full Professor
- Associate Professor
- Assistant Professor
- Post-Doc
- Doctoral Student
- Other position at a University (with follow-up question to state title)
- Outside of Academia (with follow-up question to state title)

Q40: Please describe your current position in more detail: (Textfield) (*Only displayed if answered “Other position at a University” or “Outside of Academia” to previous question*)

Q41: We are very interested to know any thoughts and comments you have about the survey you just completed or the present project to crowdsource the analysis of data. Please describe them here: (Text field)

Thank you for your answers!

Once you click submit you will be taken to the data analysis platform. Please login with your account details and begin your analysis to test the two hypotheses of interest.

Supplement 6: Post-survey for analysts in primary study

This questionnaire will be used to collect answers detailing the statistical approach that you have taken. Your answers will then be used to facilitate the online peer feedback process. Please provide enough information for a naive empiricist to be able to give you valuable feedback. Remember, not all individuals involved in this project come from the same discipline, so some methods might be unfamiliar/have a different name to those in other areas. There are two sections: one that will be shared with other researchers, and one that we will use internally to get a good first idea about actual results. Only the analytic methods will be shared with the crowdsourcing analysts to avoid bias.

Q1: What is your name?

Q2: What transformations (if any) were applied to the variables. Please be specific and explain why you applied them.

Q3 Were any cases excluded, and why?

Q4 How did you operationalize verbosity?

Q5 What are the theoretical reasons for operationalizing verbosity in that manner?

Q6 How did you operationalize status?

Q7 What are the theoretical reasons for operationalizing status in that manner?

Q8 What is the name of the statistical technique that you employed?

Q9 Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known.

Q10 Please explain why you chose this technique.

Q11 What are some references for the statistical technique that you chose?

Q12 What variables were included as covariates (or control variables) when testing Hypothesis 1: A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion.

Q13 What theoretical and/or statistical rationale was used for your choice of covariates included in the models when testing Hypothesis 1?

Q14 What variables were included as covariates (or control variables) when testing Hypothesis 2: Higher status participants are more verbose than are lower status participants?

Q15 What theoretical and/or statistical rationale was used for your choice of covariates included in the models when testing Hypothesis 2?

Q16 What unit is your effect size in?

Q17 What is the size of the effect for Hypothesis 1: A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion. Please specify the magnitude and direction of the effect size, along with the 95% confidence (or credible) interval in the following format: estimate [low interval, high interval]. Remember that this result will not be shared with other analysts at this stage.

- estimate (1)
- low interval (2)
- high interval (3)

Q18 Anything else you'd like to add?

Q19 What is the size of the effect for Hypothesis 2: Higher status participants are more verbose than are lower status participants? Please specify the magnitude and direction of the effect size, along with the 95% confidence (or credible) interval in the following format: estimate [low interval, high interval]. Remember that this result will not be shared with other analysts at this stage.

- estimate (1)
- low interval (2)
- high interval (3)

Q20 Anything else you'd like to add?

Q21 What other steps/analyses did you run that are worth mentioning? Include effect sizes in a similar format as above if necessary.

Q22 You may use the space below to paste the script you used to run the analyses. (Optional)

Q23 What is your current opinion regarding Hypothesis 1: A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion

- Very Unlikely (1)
- Unlikely (2)
- Neither Likely nor Unlikely (3)
- Likely (4)
- Very Likely (5)

Q24 Please explain why you think so.

Q25 What is your current opinion regarding Hypothesis 2: Higher status participants are more verbose than are lower status participants?

- Very Unlikely (1)
- Unlikely (2)
- Neither Likely nor Unlikely (3)
- Likely (4)
- Very Likely (5)

Q26 Please explain why you think so.

Q27 Please use this space for any additional comment you may have at this stage (this is for our information and will not displayed to others).

Please press the submit button only once you are sure that you would like to submit your responses and that no changes are needed at this stage.

Supplement 7: Quality checks of crowdsourced data analyses

Table S7-1: *Direction and significance ($\alpha = 0.05$, two-tailed) for results from the independent analysts for Hypothesis 1 and Hypothesis 2, including analyses flagged by independent statisticians as using problematic approaches.*

Hypothesis	Significant in predicted (+) direction	Not significant in predicted (+) direction)	Not significant in opposite (-) direction)	Significant in opposite (-) direction)
<i>H1: A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion</i>	50% (n=9)	11.1% (n=2)	22.2% (n=4)	16.7% (n=3)
<i>H2: Higher status participants are more verbose than lower status participants</i>	33.3% (n=5)	20% (n=3)	26.7% (n=4)	20% (n=3)

Note. For Hypothesis 1, Analyst 15 found a non-directional effect with unknown significance (McFadden's logistic regression's r-squared). For Hypothesis 2, Analyst 15 found a non-directional effect with unknown significance (McFadden's logistic regression's r-squared) and Analyst 21 found a non-directional, significant effect (eta squared). Only analyses for which both direction and significance levels are known are included in this table.

CLASSIFYING THE CODE AND RESULTS ANALYSTS SUBMITTED INTO FIVE CATEGORIES

As part of this project, we received a larger number of submissions than reported in the main text. Although many analysts submitted results and code that allowed us to understand both their analytical results and choices, other analysts submitted code, analyses, or reports of their results that lacked crucial information. The main text reports only analyses associated with detailed code and results, which the project coordination team could independently reproduce in full. All analyses, regardless of whether they met these criteria, are reported here in the supplements.

In this section, we explain the criteria according to which submitted analyses were included in the main manuscript or in the supplements.

To categorize analyses based on their completeness, we set up the coding system described below. Analysis quality was ranked from complete (5) to missing critical elements (1). We included analyses at levels 4 and 5 in the main manuscript and include results from other analyses (levels 1-3) in the supplements.

The specific criteria are provided below and in Table S7-2

- Level 5: The effect size type, estimate, degrees of freedom, p -value and screenshots of the effect size from re-running the analysis are available - we have as much complete information as possible.
- Level 4: The effect size type, estimate, and screenshots of the effect size from re-running the analysis are available, but information on either degrees of freedom or p -value is missing.
- Level 3: The effect size type, estimate, and screenshots of the effect size from re-running the analysis are available, but information on both degrees of freedom and p -value is missing.
- Level 2: Analyst has reported the effect size (ES) but the effect size could not be produced by rerunning the analysis with the code submitted by analyst.
- Level 1: Analyst has not reported any effect size (ES). No information on the analysis outcome is available.

Please refer to the following table for an overview of how the submitted analyses were distributed across these five categories.

Table S7-2: *Quality criteria and number of submitted analyses in each category.*

Quality level (and number of analyses across both hypotheses)	Included in:	Effect size available?	Effect size reproducible?	Estimates available?	Degrees of freedom available?	P values available?
5 (n= 23)	Main paper	Yes	Yes	Yes	Yes	Yes
4 (n= 13)	Main paper	Yes	Yes	Yes	Either df or <i>p</i> -values	Either df or <i>p</i> -values
3 (n= 6)	Supplement 7	Yes	Yes	Yes	No	No
2 (n= 33)	Supplement 7	Yes	No	No	No	No
1 (n= 14)	Supplement 7	No	No	No	No	No

Note. Of the n=36 analyses that satisfied all criteria for quality levels 5 (n= 23) and 4 (n= 13), 7 analyses were further identified by a sub-team of independent statisticians as containing clear errors. An in-depth review, summary, and transcription of the code for these analyses, along with detailed explanation for why they were classified as problematic can be found here: <https://osf.io/n5q3c/>

Table S7-3.1: *Direction and significance ($\alpha = 0.05$, two-tailed) for results from the independent analysts for Hypothesis 1, including analyses flagged by independent statisticians as using problematic approaches (Hypothesis 1: “A woman’s tendency to participate actively in the conversation correlates positively with the number of females in the discussion”)*

Quality level (and number of analyses for H1)	(+ effect in predicted direction			(-) effect in opposite direction			Significant, non- directional	Not significant, non- directional	Significance and direction unknown
	Significant	Significance unknown	Not significant	Significant	Significance unknown	Not significant			
5 (n= 12)	5		1	3		2			1
4 (n= 7)	4		1			2			
3 (n= 3)		1			1				1
2 (n= 15)	4	1	3		1	3	1		2
1 (n= 7)									7

Note. The table contains all analyses and results submitted, including the n= 5 Hypothesis 1 analyses at quality levels 4 and 5 that were flagged by independent statisticians as containing problematic errors. An in-depth review, summary, and transcription of the code for all analyses at levels 4 and 5 can be found here: <https://osf.io/n5q3c/>

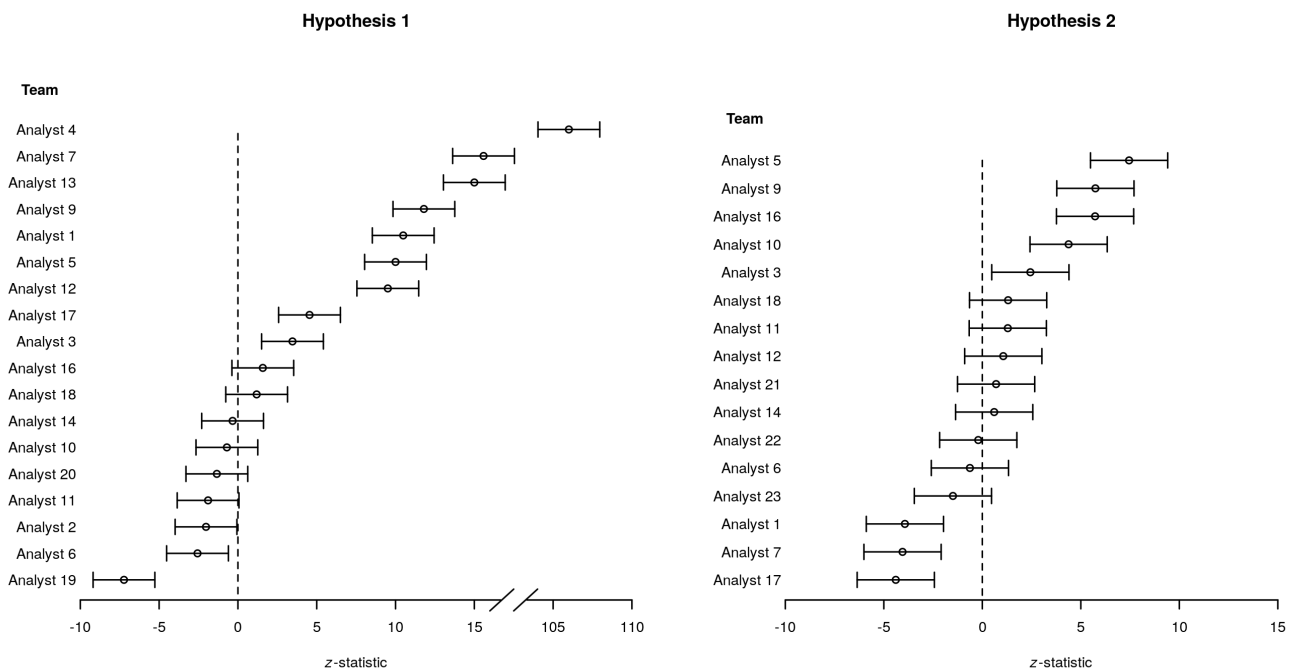
Table S7-3.2: *Direction and significance ($\alpha = 0.05$, two-tailed) for results from the independent analysts for Hypothesis 2, including analyses flagged by independent statisticians as using problematic approaches (Hypothesis 2: “Higher status participants are more verbose than lower status participants”).*

Quality level (and number of analyses for H2)	(+ effect in predicted direction			(-) effect in opposite direction			Significant, non- directional	Not significant, non- directional	Significance and direction unknown
	Significant	Significance unknown	Not significant	Significant	Significance unknown	Not significant			
5 (n= 11)	3		1	2		3		1	1
4 (n= 6)	2		2	1		1			
3 (n= 3)		2			1				
2 (n= 18)	5	1	4	1	1			1	5
1 (n= 7)									7

Note. The table contains all analyses and results submitted, including the n= 2 Hypothesis 2 analyses at quality levels 4 and 5 that were flagged by independent statisticians as containing problematic errors. An in-depth review, summary, and transcription of the code for all analyses at levels 4 and 5 can be found here: <https://osf.io/n5q3c/>

Figure S7-1 shows the standardized scores including the results of the analyses that included clear errors. In this larger set of analyses, the z -scores corresponding to the results for Hypothesis 1 ranged from -7.230 to 106.267, with a median of 2.521, and mean of 9.630 (standard error = 0.236) that was significantly different from zero ($z = 40.856$, two-tailed $p < .001$). The z -scores corresponding to the results for Hypothesis 2 ranged from -4.394 to 7.450, with a median of 0.882, and mean of 1.000 (standard error = 0.25), which was significantly different from zero ($z = 4.000$, two-tailed $p < .001$). The standardized scores were heterogeneous for both Hypothesis 1 ($\chi^2(17) = 10,636.61$, $p < .001$) and Hypothesis 2 ($\chi^2(15) = 189.54$, $p < .001$). To summarize, including the results of the analyses that contained clear errors does not affect our conclusions; both average standardized scores and conclusions based on these estimates were heterogeneous for both hypotheses.

Figure S7-1. Dispersion of z -scores corresponding to estimates of independent analysts for each hypothesis, including analyses containing problematic approaches. Note that there is a break in the x -axis of the figure for Hypothesis 1 to incorporate the extreme z -score of Analyst 4.



Supplement 8: Error checks of crowdsourced data analyses

CODING CHECKS

We conducted several checks to ensure that the dispersion of empirical results observed is not an artifact of different coding schemes for key variables. For example, Analyst A may report a positive (+) effect and Analyst B might report a negative (–) effect for the effect of status on verbosity. If Analyst A coded status such that (low status=0, high status=1) and Analyst B coded status such that (high status=0, low status=1), both analysts actually found evidence in the same direction even though their submitted effects are positive (+) and negative (–) in sign.

To ensure such coding differences in key variables do not explain the dispersion of standardized scores observed, three independent reviewers assessed the logical chain of coding of key variables for all analyses considered for the main manuscript.

Each reviewer assessed across both hypotheses:

1. How each analyst operationalized the independent variable:
 - Was the independent variable an original variable from the Edge dataset?
 - Or did the analyst create a new variable?
2. How each analyst coded the independent variable:
 - For H1:
 - + Positive such that high values = more female participation in discussion, low values = less female participation in discussion
 - - Negative such that high values = LESS female participation in discussion, low values = MORE female participation in discussion
 - For H2:
 - + Positive such that high values = high status, low values = low status
 - - Negative such that high values = LOW status, low values = HIGH status
3. How each analyst operationalized the dependent variable
 - Was the independent variable an original variable from the Edge dataset?
 - Or did the analyst create a new variable?
4. How each analyst coded the dependent variable:
 - For H1:
 - + Positive such that high values = more female tendency to participate actively, low values = less female tendency to participate actively
 - - Negative such that high values = LESS female tendency to participate actively, low values = MORE female tendency to participate actively
5. The actual result submitted by each analyst (effect size and significance level, if applicable). For example: “-1.315 (regression coefficient) , $p = 0.0429$ ” or “0.151 (logistic regression coefficient), $p < .001$ ”
6. The direction of the result, without any interpretation. For example: “- negative”, or “+ positive)

7. Conclusion regarding the direction of the effect, taking into account the coding of the independent variable, the coding of the dependent variable, and the direction of the result. This leads to a conclusion of whether the reported effect is
 - + and in the direction predicted by the hypothesis, or
 - – and in the direction contrary to the hypothesis

The table below details the complete logical chain leading towards the conclusion regarding the direction of the effect.

Table S8-1.1: *Overview of direction of effects and coding schemes for key variables for Hypothesis 1, “A woman’s tendency to participate actively in the conversation correlates positively with the number of females in the discussion”*

Analyst	Independent variable (IV) operationalization	IV coding	Dependent variable (DV) operationalization	DV coding	Result	Direction of effect
1	sum of previous female contributors	+	whether next contributor in thread is female	+	1.063 (odds ratio from logistic regression model), $p < 0.001$	+
2	number of female contributors organized by order in which posted in each conversation	+	total number of posts each female contributor makes in each conversation	+	-1.3152 (regression coefficient), $p = 0.043$	-
3	"UniqueFemaleParticipation" (x10)	+	"ContributionsbyAuthor" (only females)	+	0.333 (regression coefficient), $p = .0005$	+
4	“UniqueFemaleContributors”	+	"FemaleContributions"	+	0.870 (Pearson’s correlation coefficient), $p < .001$	+
5	"FemaleParticipation"	+	"FemaleContributions"	+	0.559 (Pearson’s correlation coefficient), $p < .001$	+
6	“UniqueFemaleContributors”	+	change in number of contributions made by each participant in the past versus now (only females)	+	-0.593 (regression coefficient), $p = .0103$	-
7	sum of previous female contributions	+	"Female"	+	0.150 (logistic regression coefficient), $p < .001$	+
9	cumulative proportion of females in each thread	+	"WC" (only females)	+	23.467 (regression coefficient), $p < .001$	+

11	"UniqueFemaleContributors" (modified)	+	"ContributionsbyAuthor" (only females)	+	-0.0233 (regression coefficient), $p = .0583$	-
12	"UniqueFemaleContributors"/"UniqueContributors"	+	"FemaleContributions"	+	27.3 (incidence rate ratio = regression coefficient for Poisson regression raised to exponential form), $p < .001$	+
13	"UniqueFemales" (modified)	+	"FemaleParticipation" (modified)	+	0.259 (regression coefficient), $p < .001$	+
14	"UniqueFemaleContributors"	+	"FemaleParticipation"	+	-0.001 (regression coefficient), $p = .736$	-
17	"UniqueFemaleParticipation"	+	"Female_Contributions"/ "UniqueFemaleContributors"	+	0.368 (correlation coefficient), $p < .001$	+
19	"UniqueFemaleContributors"	+	"ContributionsbyAuthor" (only females)	+	-0.3155 (regression coefficient), $p < .001$	-

Note. Variable names from the original dataset are written within "" (example: "UniqueFemaleParticipation"). Variables not from the original dataset are described briefly.

A "+" coding of the independent variable indicates that higher values in the independent variable represent more females in the discussion and lower values in the independent variable represent fewer females in the discussion.

A "+" coding of the dependent variable indicates that higher values in the dependent variable represent a woman's greater tendency to participate actively in the conversation and low values in the dependent variable represent a woman's lower tendency to participate actively in the conversation.

Table S8-1.2: *Overview of direction of effects and coding schemes for key variables for Hypothesis 2, “Higher status participants are more verbose than lower status participants”*

Analyst	Independent variable (IV) operationalization	IV coding	Dependent variable (DV) operationalization	DV coding	Result	Direction of effect
1	"AcademicHierarchyStrict"	+	log of "Number_Characters"	+	-0.162 (log-level linear regression), $p < .001$	-
3	"Total_Citations" (scaled)	+	"WC" (scaled)	+	.0434 (regression coefficient), $p = .0194$	+
5	“Job_Title_S” (chaired professor with baseline: assistant professor)	+	“ThreadsThisYear”	+	3.97 (regression coefficient), $p < .001$	+
6	"AcademicHierarchyStrict"	+	average number of words for each participant / thread	+	-64.38 (regression coefficient), $p = .53$	-
7	base-10 log of "Citations_Cumulative"	+	base-10 log of "Number_Characters"	+	-0.221 (regression coefficient), $p < .001$	-
9	index of "Workplace_SR_Bin", "HavePhD"	+	mean of "WC"	+	69.70 (regression coefficient), $p < .001$	+

10	index of seven status-related variables (five positively coded: "Academic, PhD_Institution_US_IR_Bin, threadPart, Total_Citations, AcademicHierarchyStrict", two negatively coded: "HavePhD, Workplace_US_IR_Bin")	+	log of "WC"	+	0.122 (regression coefficient), $p < .001$	+
11	"H_Index"	+	"WPS"	+	.090 (regression coefficient), $p = .196$	+
12	"AcademicHierarchyStrict"	+	average number of words per participant	+	54.39 (regression coefficient), $p = .2874$	+
14	"Workplace_US_Bin"	-	log of "Number_Characters"	+	0.059 (regression coefficient), $p = .549$	-
17	"AcademicHierarchyStrict"	+	"WC"	+	-0.053 (Kendall's Tau correlation coefficient), $p < .001$	-
18	index of "PhD_Year" "AcademicHierarchyStrict" through PCA	+	index of "FracCharSpoke", "FracTimesSpoke" through PCA	+	0.133 (regression coefficient), $p = .192$	+

21	“AcademicHierarchyStrict” (modified)	+	average number of words for each unique user	+	.019 (eta squared), $p = .242$	non-directional
22	“AcademicHierarchyStrict” (modified: 2013)	+	number of contributions for each participant (2013)	+	-0.037 (Spearman's correlation coefficient), $p = .582$	-
23	“AcademicHierarchyStrict” (modified: characters)	+	average number of characters made by each participant for each level of “AcademicHierarchyStrict”	+	-239.01 (regression coefficient), $p = .136$	-

Note. Variable names from the original dataset are written within “” (example: “UniqueFemaleParticipation”). Variables not from the original dataset are described briefly.

A “+” (-) coding of the independent variable indicates that higher (lower) values in the independent variable represent high level of status and lower (higher) values in the independent variable represent low levels of status.

A “+” coding of the dependent variable indicates that higher values in the dependent variable represent high levels of verbosity and low values represent low levels of verbosity.

CODING CHECKS (PILOT PHASE)

In addition, we also conducted coding checks for the 12 analyses submitted during the pilot phase of the project (see Supplement 3). For these analyses, three independent reviewers checked whether the observed dispersion in standardized scores could be explained by differences in the analysts' coding scheme as described above. This was not the case.

INDEPENDENT REANALYSIS OF SUBMITTED RESULTS

We also wanted to make ensure that the results reported by analysts were fully reproducible from the code the analysts had submitted. We wanted to exclude the possibility that an analyst accidentally submitted a different result than that produced by their R code.

For example, Analyst A may have seen a correlation coefficient of “.389” in R. However, the analyst may have accidentally submitted a correlation coefficient different from the one that was displayed, or have flipped the sign of the displayed effect (for example, Analyst A could have possibly submitted another value such as -.389, or .983 or .398).

To further increase confidence in the dispersion of observed in standardized scores, we sought to independently replicate the results reported by analysts by doing the following for each of the individual results submitted by analysts:

- Open the raw dataset
- Apply the code submitted by the analysts step-by-step to the raw code
- Insert comments as the code is run to better conceptually understand and capture the logic of the analyst's code
- Identify which of the empirical models included in the code produced the final result submitted by the analyst

As an additional accuracy check we took screenshots of each of these reproduced results. This allowed us to confirm for each of the submitted results that they were actually the product of the code submitted by the analyst and not the result of some potential error or reporting issue.

DETAILED SUMMARY AND TRANSCRIPTION OF EACH ANALYST'S CODE

The R codes submitted by analysts can be quite dense, and readers who wish to understand each analyst's statistical choices and reasoning may have to invest some time to do so. To help the reader quickly and conveniently understand the details of each analyst's code, three reviewers independently summarized and transcribed each analyst's code.

This provides the reader with a line-by-line summary and interpretation of each analyst's code. To maximize convenience for the reader we compiled this information into one table for each of the submitted analyses. Each table contains the following columns:

#Ln	Main task	Var name	Code description
describes the lines that form a meaningful code block	main task a code block had to achieve in operational terms This could range from “ <i>creating a new variable</i> ” to “ <i>running linear regression model</i> ” and “ <i>exploring model assumptions are met</i> ” to “ <i>computing Bayes factors</i> ”	identifies the variables the analyst is working on at each stage	a meaningful summary and interpretation of the analyst's intention behind a code block

For example, below is a short segment of this for Analyst 21. Note that the full code for this analyst contains 603 lines.

In #	Main Task	Var Name	Code Description
2-5	Set up file	-	Set up the RMD (R Mark Down) file
10	Clean Environment		Clear the workspace and delete all previous variables
22-31	Load packages	-	Load the packages: (tidyverse, BayesFactor, psych, ccaPP, BayesMed, rjags, dummies, ez, heplots, MBESS)
33	Read the data	d	Read the CSV file “edge1.1”
49	Explore data	d\$year	Plot a barplot for the frequency of the variable “year” while setting the y limit between (150 and 1500)
56			Create a contingency table for the variables “Year” and “Type” from the dataframe (d)
69	Plot data		Divide the display into two parts
70			Plot a barplot for the frequency of the variable “Male_Contributions” with the graph title (male) from the dataframe (d)

Each summary also contains the raw code so the reader can access both the original raw code and the summarized version in one convenient place. Each summary also contains two additional error checks as described below.

The complete summaries can be accessed on the Open Science Framework (OSF) here: <https://osf.io/n5q3c/>.

IDENTIFYING THE ANALYTICAL COMPONENTS OF EACH ANALYSIS

To further help the reader assess the submitted results, we also identified the key analytical components for each analysis and present these in a convenient way to help the reader navigate each code. Specifically, for each submitted result we extracted the following information in one convenient location:

0. Sample size used in final empirical model and analysis:
 - Split up by male/female if applicable
1. Unit of analysis:
 - For example: comment, thread, conversation, etc.
2. Verbal interpretation of the result:
 - A verbal summary of how the result can be interpreted
3. Data Filters:
 - Explanation of whether the analyst decided to include only certain cases in the analysis
4. Dependent variable operationalization:
 - Description of how the analyst decided to operationalize the dependent variable. In many cases, these were taken directly from the original dataset without any further modifications. Other analysts transformed original dataset variables (log-transformed or some other transformation), or built completely new variables.
5. Independent variable operationalization
 - Same as dependent variable operationalization above, but applied to the IV

Below we provide a sample entry for the reader's convenience. This information can be accessed for all submitted analyses – along with the summaries and error checks for each analysis on the OSF here: <https://osf.io/n5q3c/>

Number of observations used in final empirical model and analysis: n = 4262 (7975 before missing values are dropped)

*# males = 3855 (based on variable *Female*)*

*# females = 407 (based on variable *Female*)*

“Missing” = 3713

Unit of measurement: Contributions (i.e. one observation for every contribution made in the form of a post or comment)

1. High level statement of the analysis:

The analyst reported an effect size of -0.16150 (based on a log-level linear regression model). The p-value reported is < 0.001.

2. Verbal interpretation of the result:

On average, an increase in *AcademicHierarchyStrict* by 1 unit, decreases *Number.Characters* by -16.15% (equivalent to a -0.16150 decrease in *log_num_char*) after controlling for *Female* and *Academic*.

3. Data filters:

None used. However, 3713 observations were dropped during the regression analysis “due to missingness.” These observations were dropped automatically by the regression because there were missing values (NA) in one or more of the regression variables involved.

4. Dependent variable operationalization:

log_num_char is simply the natural log of *Number.Characters* which is the number of characters in a given entry.

5. Independent variables operationalization:

The independent variables are *AcademicHierarchyStrict*, *Female*, and *Academic*. All variables are as-is from the original Edge data set. *AcademicHierarchyStrict* was used as the explanatory variables associated with Hypothesis 2.

ERROR CHECKS

We established that the dispersion of standardized scores observed across analysts is not explained by differential coding schemes of key variables (see above). In addition, we wanted to ensure that the observed dispersion of standardized scores is not explained by clear analytical errors throwing off estimates. We therefore also conducted thorough checks for errors in each analysis.

To be conservative, we only include analyses in the main paper which – to the best of our knowledge – do not contain clear errors. We also present the results including the results from all eight analyses that our team of independent statisticians found to contain clear errors, in Supplement 7.

These error checks were conducted by three independent coders who examined each submitted analysis in dyads of two for errors. Clear errors were described as fundamental analytical errors that might threaten the validity of the analysis. The three coders were instructed to highlight any such potential errors, and – if in doubt – to flag an analysis as potentially containing an error rather than not flagging it. When one of the two coders making up a dyad flagged an analysis as potentially containing errors, the dyad reviewed and discussed the potential reasons for doing so. The dyad then either declared an analysis as containing clear errors or not.

If no agreement was reached or if any kind of uncertainty remained the two independent coders making up one dyad consulted with a third coder. Analyses for which any kind of uncertainty remained about the presence of errors were further discussed by a team of statistical experts.

This process resulted in 7 analyses that were flagged as containing clear errors, and are therefore not discussed or presented in the main paper. Details on these analyses can be found in Supplement 7.

We have pasted an example below. The information on errors can be accessed for all submitted analyses, along with the summaries and key analytical components for each analysis, on the OSF here: <https://osf.io/n5q3c/>.

Hypothesis: 2 Analyst 1 <u>Reviewer:</u> Reviewer 1 <u>Co-reviewer:</u> Reviewer 3											
<p align="center"><u>Error check 1: Gross analytical errors?</u></p> <p><i>Did you detect any gross or obvious errors in the final empirical model submitted by the analyst to test this hypothesis</i></p> <table border="1"> <thead> <tr> <th></th> <th><i>No, I did not detect any gross or obvious errors in the final empirical model.</i></th> <th><i>Yes, I did detect any gross or obvious errors in the final empirical model</i></th> </tr> </thead> <tbody> <tr> <td><i>Reviewer</i></td> <td align="center">X</td> <td></td> </tr> <tr> <td><i>Co-reviewer</i></td> <td align="center">X</td> <td></td> </tr> </tbody> </table> <p><i>Main Reviewer, do you have any additional comments?</i></p> <ul style="list-style-type: none"> The analyst uses <i>AcademicHierarchyStrict</i> as a proxy for the status of participants while controlling for <i>Academic</i>. <p><i>Co reviewer, do you have any additional comments?</i></p> <ul style="list-style-type: none"> No 				<i>No, I did not detect any gross or obvious errors in the final empirical model.</i>	<i>Yes, I did detect any gross or obvious errors in the final empirical model</i>	<i>Reviewer</i>	X		<i>Co-reviewer</i>	X	
	<i>No, I did not detect any gross or obvious errors in the final empirical model.</i>	<i>Yes, I did detect any gross or obvious errors in the final empirical model</i>									
<i>Reviewer</i>	X										
<i>Co-reviewer</i>	X										

ERROR CHECKS (PILOT PHASE)

The same 3 independent reviewers who checked the 12 analyses submitted during the pilot phase of the project for coding differences also checked whether errors might explain the dispersion in standardized scores observed. This was not the case.

HOLISTIC JUDGMENT OF SUBMITTED RESULTS

In addition to these checks, we also examined the code and results submitted by analysts holistically. To do this, we examined verbal commentary contained in the code or the submitted results. We did this to assess whether any of these responses might indicate an analyst's interpretation of the results was not actually in line with the results that she or he submitted.

Careful inspection of each these verbal responses indicated this was not the case and analysts were conceptually aligned with their submitted statistical results.

APPENDIX S8-1

The reader can access the error checks, key analytical components, and a detailed summary of each submitted analysis on the OSF here: <https://osf.io/n5q3c/>. Below is one sample document for one submitted analysis containing these three components.

Hypothesis: 2

Analyst 21

Reviewer: Reviewer 3

Co-reviewer: Reviewer 1

Error check 1: Gross analytical errors?

Did you detect any gross or obvious errors in the final empirical model submitted by the analyst to test this hypothesis

	<i>No, I did not detect any gross or obvious errors in the final empirical model.</i>	<i>Yes, I did detect any gross or obvious errors in the final empirical model</i>
<i>Reviewer</i>	X	
<i>Co-reviewer</i>	x	

Main Reviewer, do you have any additional comments?

- The code includes analysis for both analysis 1 and 2. In this review I processed the code that is specific for H2 only.

Co reviewer, do you have any additional comments?

- I believe that the eta-squared value for an ANOVA test is always positive. It only captures how different the means of the dependent variable are, but not in which direction.
- While I would not classify this as a gross error, not having a direction for the results may make it difficult to interpret the results.

Error check 2: Identifying analytical components of the final model

Sample size used in final empirical model and analysis: n = 355

Unit of analysis: Participants

Females = 49

Males = 306

1. High level statement of the analysis:

Analyst 21 has reported an eta squared value= 0.01899725 and p-value= 0.2421 (Insignificant).

2. Verbal interpretation of the result:

1.8% of the variance of “ContrTotal” is explained by “AcH2” while removing the effects of other sources.

3. Data filters:

The analyst has used the edge dataset with no filters, but he/she did remove the missing values from the data.

4. Dependent variable operationalization:

The dependent variable “ContrTotal” is not an original variable from the Edge dataset. It captures the average number of words for each unique user (Id). Details on how this variable was constructed can be found in the code summary below.

5. Independent variables operationalization:

The independent variable “AcH2” is not an original variable from the Edge dataset. It is the value of “AcademicHierarchyStrict” for each participant but with a labeled hierarchy.

Both the dependent and independent variables are taken from the dataset (dd2), for specific details on how the dataset has been created, please check line #578-587 in the following section.

Verbal summary of code submitted by analyst:

In #	Main Task	Var Name	Code Description
2-5	Set up file	-	Set up the RMD (R Mark Down) file
10	Clean Environment		Clear the workspace and delete all previous variables
22-31	Load packages	-	Load the packages: (tidyverse, BayesFactor, psych, ccaPP, BayesMed, rjags, dummies, ez, heplots, MBESS)
33	Read the data	d	Read the CSV file “edge1.1”
49	Explore data	d\$year	Plot a barplot for the frequency of the variable “year” while setting the y limit between (150 and 1500)
56			Create a contingency table for the variables “Year” and “Type” from the dataframe (d)
69	Plot data		Divide the display into two parts
70			Plot a barplot for the frequency of the variable “Male_Contributions” with the graph title (male) from the dataframe (d)
71			Plot a barplot for the frequency of the variable “Female_Contributions” with the graph title (female) from the dataframe (d)
75	Explore data	d\$Male_Contributions	Create a frequency table for the variable “Male_Contributions” from the dataframe (d)
76		d\$Female_Contributions	Create a frequency table for the variable “Female_Contributions” from the dataframe (d)
78		d\$Male_Contributions	Create the frequency table for the variable “Male_Contributions”. Transform the first row of the table into numeric variable and return its sum. (Return the sum of “Male_Contributions”) from the dataframe (d)
79		d\$Female_Contributions	Create the frequency table for the variable “Female_Contributions”. Transform the first row of the table into numeric variable and return its sum. (Return the sum of “Female_Contributions”) from the dataframe (d)

85	Plot data		Plot a barplot for the frequency of the variable “FemaleParticipation” from the dataframe (d)
88			Plot a barplot for the frequency of the variable “NumberOfAuthorContributions” from the dataframe (d)
91			Plot a barplot for the frequency of the variable “NumberOfAuthorContributions” but only considering entries where “NumberOfAuthorContributions” > 0 from the dataframe (d)
94			Plot a barplot for the frequency of the variable “DebateSize” from the dataframe (d)
100			Plot a barplot for the frequency of the variable “UniqueContributors” with the graph title (Unique Contributors) from the dataframe (d)
101			Plot a barplot for the frequency of the variable “UniqueMaleContributors” with the graph title (Unique Male) from the dataframe (d)
102			Plot a barplot for the frequency of the variable “UniqueFemaleContributors” with the graph title (Unique Female) from the dataframe (d)
104			Plot a barplot for the frequency of the variable “UniqueFemaleParticipation” with the graph title (Unique Female Participation) from the dataframe (d)
111			Divide the display into two parts
112			Plot a pie chart for the variable “Male” with the graph title (Male) from the dataframe (d)
113			Plot a pie chart for the variable “Female” with the graph title (Female) from the dataframe (d)
118	Explore data	d\$Job_Title_S	Create a frequency table for the variable “Job_Title_S” from the dataframe (d)
122		d\$ContributionsThisYear	Create a frequency table for the variable “ContributionsThisYear” from the dataframe (d)

124		d\$ThreadsThisYear	Create a frequency table for the variable “ThreadsThisYear” from the dataframe (d)
128		d\$PhD_Institution_US_IR	Create a frequency table for the variable “PhD_Institution_US_IR” from the dataframe (d)
132		d\$PhD_Institution_US	Create a frequency table for the variable “PhD_Institution_US” from the dataframe (d)”
138	Plot data		Plot a barplot for the frequency of the variable “H_Index” from the dataframe (d)
142			Plot a barplot for the frequency of the variable “i10_Index” from the dataframe (d)
146			Plot a barplot for the frequency of the variable “Citations_Cumulative” from the dataframe (d)
150			Plot a barplot for the frequency of the variable “Number.Characters” from the dataframe (d)
151	Explore data	d\$Number.Characters	Print the descriptive statistics of the variable “Number.Characters” from the dataframe (d)
159	Plot data		Generate a density plot for the variable “Female_Contributions” from the dataframe (d)
161			Generate a density plot for the variable “UniqueFemaleContributors” from the dataframe (d)
165	Test Normality	d\$Female_Contributions	Check normality of the variable “Female_Contributions” from the dataframe (d) using Kolmogorov Smirnov test. (The variable is not normally distributed)
166		d\$UniqueFemaleContributors	Check normality of the variable “UniqueFemaleContributors” from the dataframe (d) using Kolmogorov Smirnov test. (The variable is not normally distributed)
175	Plot data		Generate a scatter plot for “UniqueFemaleContributors” on the y-axis against “Female_Contributions” on the x-axis from the dataframe (d)
180	Remove zeros and plot again	dtmp	Select the two columns (Female_Contributions, UniqueFemaleContributors) from (d) and put them in a new dataframe (dtmp)

181			Remove the rows where both (Female_Contributions, UniqueFemaleContributors) are zeros in (dtmp)
182			Generate a scatter plot for “UniqueFemaleContributors” on the y-axis against “Female_Contributions” on the x-axis from the dataframe (dtmp)
196	Plot data		Generate a scatter plot for “UniqueFemaleParticipation” on the y-axis against “FemaleParticipation” on the x-axis from the dataframe (d)
197			Generate a scatter plot for “FemaleParticipation” on the y-axis against “UniqueFemaleParticipation” on the x-axis from the dataframe (d)
205	Test Correlation		Test Kendall correlation between (UniqueFemaleParticipation, FemaleParticipation) in the dataframe (d) while returning a consistent estimate for the normal distribution.
206			Test Kendall correlation between (UniqueFemaleParticipation, FemaleParticipation) in the dataframe (d).
211	Explore data	d\$UniqueFemaleParticipation	Return the descriptive statistics of the variable “UniqueFemaleParticipation” from the dataframe (d)
212		d\$FemaleParticipation	Return the descriptive statistics of the variable “FemaleParticipation” from the dataframe (d)
214	Check & Remove Outliers	d\$UniqueFemaleParticipation	Remove the values of “UniqueFemaleParticipation” from the dataframe (d) which are bigger than: [mean(UniqueFemaleParticipation) + 3 standard deviations].
216		d\$FemaleParticipation	Remove the values of “FemaleParticipation” from the dataframe (d) which are bigger than: [mean(FemaleParticipation) + 3 standard deviations].
218		d\$UniqueFemaleParticipation	Remove the values of “UniqueFemaleParticipation” from the dataframe (d) which are smaller than: [mean(UniqueFemaleParticipation) - 3 standard deviations].
220		d\$FemaleParticipation	Remove the values of “FemaleParticipation” from the dataframe (d) which are smaller than: [mean(FemaleParticipation) - 3 standard deviations].

224		d1	Create new dataframe (d1) which equals the dataframe (d) after removing the values of “UniqueFemaleParticipation” which are bigger than: [mean(UniqueFemaleParticipation) + 3 standard deviations].
225		d1	Modify the dataframe (d1) by removing also he values of “FemaleParticipation” which are bigger than: [mean(FemaleParticipation) + 3 standard deviations].
227	Plot data		Generate a scatter plot for “FemaleParticipation” on the y-axis against “UniqueFemaleParticipation” on the x-axis from the dataframe (d1)
229	Test Correlation		Test Kendall correlation between (UniqueFemaleParticipation, FemaleParticipation) in the dataframe (d1) while returning a consistent estimate for the normal distribution.
230			Test Kendall correlation between (UniqueFemaleParticipation, FemaleParticipation) in the dataframe (d1).
233 - 419: Testing hypothesis #2			
424	Explore data		Create a frequency table for the variable “AcademicHierarchyStrict” from the dataframe (d)
425	Test correlation		Test Kendall correlation between (AcademicHierarchyStrict & ContributionsThisYear) from the dataframe (d)
426-427	Explore data	d\$Workplace_SR	Create a frequency table for the variable “Workplace_SR” from the dataframe (d)
430		d\$Workplace_US	Create a frequency table for the variable “Workplace_US” from the dataframe (d)
431		d\$Workplace_SR	Return the sum of the frequency table of “Workplace_US”
432		d	Return the number of rows of (d)
433		d\$Id	Return the number of unique “Ids” from the dataframe (d)
435			Filter the dataframe (d) to get rows where Id = “richard_dawkins”. From the filtered rows select: (Year, PreviousContributions, ContributionsThisYear)

			Sort the rows by the “Year” descendingly Slice only the first row and sum it
438	Create new dataframe	dm	Column bind the dataframe (d) with newly created one (ContrTotal) that is defined by selecting: (PreviousContributions, ContributionsThisYear) and summing over the rows.
439	Explore data		Select the columns (Id, AcademicHierarchyStrict) from the dataframe (dm), remove duplicates, and create the frequency table for “AcademicHierarchyStrict”
440			Select the columns (Id, AcademicHierarchyStrict) from the dataframe (dm), remove duplicates, and return the number of rows.
450	Create new dataframe	dmm	Create the new dataframe from (dm) by grouping the latter by the variable “Id”, and for each group define “ContrTotal” as max(ContrTotal)
451		dmm\$Ac	Add new column to the new dataframe (dmm) named “Ac”. The new column is all “NA”
452-454			The values of the variable “Ac” is defined as follows: Loop from 1 till the number of rows of (dmm), and for each element: Select the corresponding rows from (dm) where “dm\$Id” = (“dmm\$Id[i]”, the “Id” that is being looped over). From these rows, select the first value of the variable “AcademicHierarchyStrict”, unlist it, and transform it to numeric variable. (For each Id, get the corresponding value of “AcademicHierarchyStrict”)
456		dmmf	Create a new dataframe (dmmf) which = (dmm) after removing the NA values.
457	Create new dataframe	dmmf\$Ac2	Create new variable “Ac2” in the dataframe which is a sorted version of the variable “Ac”
458		dmmf\$Ac3	Create new variable “Ac3” in the dataframe which = “Ac” but transformed into a factor.
460	Compute Bayes		Compute the Bayes factor for the following ANOVA design:

	factors		“ContrTotal” is the dependent variable, “AC” is the independent variable, the variable “Id” is random, and compute it without showing the progress bar. All the variables are from the dataframe (dmmf)
461			Compute the Bayes factor for the following ANOVA design: “ContrTotal” is the dependent variable, “AC” is the independent variable, the variable “Id” is random, and compute it without showing the progress bar. Consider only the entries from the dataframe (dmmf) where “ContrTotal” < 101.
464-465	Plot data		Generate a scatter plot for “ContrTotal” on the y-axis against “Ac” on the x-axis. For both variables, consider only the entries from the dataframe (dmmf) where “ContrTotal” < 101. Add red points to the plot for the values of “Ac” where “ContrTotal” < 101. Both are from the dataframe (dmmf)
467	Modify dataframe	d	Column bind the dataframe (d) with newly created one (ContrTotal) that is defined by selecting: (PreviousContributions, ContributionsThisYear) from (d) and summing over the rows.

470 - 574: Re-analyzing hypothesis #2			
578	Modify dataframe	d	<p>[Main reviewer's comment: In order to run this line, the reviewers had to use a library the analyst did not include in the code. This does not affect the result at all, it is just to understand the variable]</p> <p>Column bind the dataframe (d) with newly created variable (nWords) that is defined as follows: (Count the number of words in each comment)</p> <ul style="list-style-type: none"> • Select (Text) from the dataframe (d). • Create a function that transforms the text from each row into a character, split it into words, unlist them, and count them. The number of words is what is returned in "nWords"
580	Create new dataframe	d2	Create the new dataframe (d2) from the dataframe (d) by grouping the latter by the variable "Id", and for each group define "ContrTotal" as mean(nWords) (Get the average number of words per Id)
581		d2\$AcH	Add new column to the new dataframe (d2) named "AcH". The new column is all "NA"
582-584			<p>The values of the variable "Ac" is defined as follows:</p> <p>Loop from 1 till the number of rows of (d2), and for each element:</p> <p>Select the corresponding rows from (d) where "d\$Id" = ("d2\$Id[i]", the "Id" that is being looped over). From these rows, select the first value of the variable "AcademicHierarchyStrict", unlist it, and transform it to numeric variable.</p> <p>(For each Id, get the corresponding value of "AcademicHierarchyStrict", if there is more than one value for "AcademicHierarchyStrict", take the first one.)</p>
586		dd2	Create a new dataframe (dd2) which = (d2) after removing the NA values.
587	Create new dataframe		Create new variable "AcH2" in the dataframe that is an exact copy of "AcH" but with a labeled hierarchy for the values.
589	Plot data		Plot the variable "AcH2" on the y-axis against "ContrTotal" on the x-axis. Both variables are unlisted and from the dataframe (dd2)

590			Generate a bar plot for the frequency of the variable “ContrTotal” from the dataframe (dd2). The names of the bars are taken from “dd2\$AcH2”
592	Test Normality	dd2\$ContrTotal	Check normality of the variable “ContrTotal” from the dataframe (dd2) using Kolmogorov Smirnov test. (It is not normally distributed)
593	Run factorial ANOVA	an	Run factorial ANOVA with “ContrTotal” as dependent variable, “AcH2” as independent variable. The identifier is the column “Id” All the data is from the dataframe (dd2) an <- ezANOVA(data = dd2, dv = ContrTotal, between = AcH2, wid = Id, return_aov = TRUE) (Factorial ANOVA compares means between two or more independent variables)
594			Return eta squared for the model (an)
595	Check confidence intervals		Return the exact confidence intervals for the proportion of variance accounted for. The used formula is: ci.pvaf(F = as.numeric(an\$ANOVA["F"]), df.1 = as.numeric(an\$ANOVA["DFn"]), df.2 = as.numeric(an\$ANOVA["DFd"]), N = nrow(dd2), conf.level = 0.95)

Raw analyst code

File: [Analyst 21].Rmd

001: ---

002: title: "Analyses Report of [Analyst 21]"

003: output:

004: html_document: default

005: html_notebook: default

006: ---

007: # -----

008: ```{r}

009: # Clear workplace

010: rm(list = ls(all = T))

011: # Install and load libraries

```

012: #install.packages("tidyverse")
013: #install.packages("BayesFactor")
014: #install.packages("psych")
015: #install.packages("ccaPP")
016: #install.packages("BayesMed")
017: #install.packages("rjags")
018: #install.packages("dummies")
019: #install.packages("ez")
020: #install.packages("heplots")
021: #install.packages("MBESS")
022: suppressPackageStartupMessages(library(tidyverse))
023: suppressPackageStartupMessages(library(BayesFactor))
024: suppressPackageStartupMessages(library(psych))
025: suppressPackageStartupMessages(library(ccaPP))
026: suppressPackageStartupMessages(library(BayesMed))
027: suppressPackageStartupMessages(library(rjags))
028: suppressPackageStartupMessages(library(dummies))
029: suppressPackageStartupMessages(library(ez))
030: suppressPackageStartupMessages(library(heplots))
031: suppressPackageStartupMessages(library(MBESS))
032: # read data
033: d <- read.csv("edge1.1.csv")
034: # Check header of data
035: # UNCOMMENT TO RUN d %>% head
036: # Check all data
037: # UNCOMMENT TO RUN d %>% View
038: # From the first look at it, it looks that the
039: # data are in good order and that the titles correspond
040: # to the titles given by the authors in the 'Variable Description.docx'
041: ```
042: # -----
043: # -----
044: ```{r}
045: # After loading the data, we are going to investigate a bit the different variables so that we
can

```

```

046: # have an idea about how the data look like. This is important to do before
047: # moving to the main analyses as we may have to do also a lot of cleaning
048: # before the main analyses -- e.g., a lot of NAs or weird variables.
049: d %>% dplyr::select(Year) %>% table %>% barplot(ylim = c(150, 1500), xpd = FALSE,
las = 0)
050: # From the barplot we see that there is variability in the number of lines per year.
051: # This could be for different reasons that will be probably be explained later as we explore
052: # the data set.
053: # -----
054: # Now we are going to check what type of conversations we have, and how those
conversations are
055: # distributed per year
056: d %>% dplyr::select(Year, Type) %>% table()
057: # Apart from years 1996, and 1997 there seems to be more contributions in converstations
than
058: # the annual questions. This makes sense of course.
059: # -----
060: # Now, what could be interesting is to see how much unique are these questions
061: # UNCOMMENT TO RUN d %>% dplyr::select(Title, Year, Type) %>% table
062: # OK, that is a bit too long to read and not informative enough
063: ```
064: # -----
065: # -----
066: ```{r}
067: # OK, maybe things get a bit interesting now. We are going to run the
068: # descriptive for male and female contributions
069: layout(1:2)
070: d %>% dplyr::select(Male_Contributions) %>% table %>% barplot(main = "male")
071: d %>% dplyr::select(Female_Contributions) %>% table %>% barplot(main = "female")
072: # So, if I am understanding this right, men speak much more than women. There
073: # is a problem with the y axis, but still, this is what it looks like. I will
074: # also run frequencieis just to see the raw numbers
075: d %>% dplyr::select(Male_Contributions) %>% table()
076: d %>% dplyr::select(Female_Contributions) %>% table()
077: # ... and the sums

```

```

078: attr(table(dplyr::select(d, Male_Contributions)), "dimnames")[[1]] %>% as.numeric()
%>% sum()

079: attr(table(dplyr::select(d, Female_Contributions)), "dimnames")[[1]] %>% as.numeric()
%>% sum()

080: # OK, huge differences here. Preliminary but still a first look on this variables
081: ```
082: # -----
083: ```{r}
084: # Seeing female participation
085: d %>% dplyr::select(FemaleParticipation) %>% table %>% barplot
086: # These are way too low percentages
087: # Number of authors
088: d %>% dplyr::select(NumberOfAuthorContributions) %>% table %>% barplot
089: # OK good. Maybe this is a bit wrong analysis because there are many zeros (this
090: # refers to the annual questions). So, once again without the zeros
091: d %>% dplyr::select(NumberOfAuthorContributions) %>%
dplyr::filter(NumberOfAuthorContributions > 0) %>%table() %>% barplot()
092: #OK good.
093: # Debate Size
094: d %>% dplyr::select(DebateSize) %>% table() %>% barplot()
095: # So, many conversations have large sizes from what I can see
096: ```
097: # -----
098: ```{r}
099: # Descriptives of unique contributors
100: d %>% dplyr::select(UniqueContributors) %>% table() %>% barplot (main = "Unique
Contributors")
101: d %>% dplyr::select(UniqueMaleContributors) %>% table() %>% barplot (main =
"Unique Male")
102: d %>% dplyr::select(UniqueFemaleContributors) %>% table %>% barplot (main =
"Unique Female")
103: # From here we see again that more males participate than females
104: d %>% dplyr::select(UniqueFemaleParticipation) %>% table() %>% barplot (main =
"Unique Female Participation")
105: # Super low percentages
106: ```
107: # -----

```

```

108: ```{r}
109: # We skip some variables and we do descriptives for males and females
110: # Here we do piecharts just because some people hate them
111: layout(t(1:2))
112: d %>% dplyr::select(Male) %>% table() %>% pie (main = "Male")
113: d %>% dplyr::select(Female) %>% table() %>% pie (main = "Female")
114: ```
115: # -----
116: ```{r}
117: # Job titles
118: d %>% dplyr::select(Job_Title_S) %>% table()
119: # BTW, we do not have data for 254 participants. Most participants are professors (2442)
120: # -----
121: # Contributions this year
122: d %>% dplyr::select(ContributionsThisYear) %>% table()
123: # Most people participate 1s.
124: d %>% dplyr::select(ThreadsThisYear) %>% table()
125: # Again, once is the main number
126: # -----
127: # PhD Institution US IR
128: d %>% dplyr::select(PhD_Institution_US_IR) %>% table()
129: # Most people come from the 1st university?
130: # -----
131: # PhD Institution US
132: d %>% dplyr::select(PhD_Institution_US) %>% table()
133: # Again 1st is the dominant
134: ```
135: # -----
136: ```{r}
137: # H_Index
138: d %>% dplyr::select(H_Index) %>% table() %>% barplot()
139: # Most people are around 50 index
140: # -----
141: # i10_index
142: d %>% dplyr::select(i10_Index) %>% table() %>% barplot()

```

```

143: # It is around 70
144: # -----
145: # Citations_Cumulative
146: d %>% dplyr::select(Citations_Cumulative) %>% table() %>% barplot()
147: # Around 5000 publications
148: # -----
149: # Number_Characters
150: d %>% dplyr::select(Number.Characters) %>% table() %>% barplot()
151: d %>% dplyr::select(Number.Characters) %>% describe()
152: ```
153: # -----
154: ```{r}
155: # This is the very basic hypothesis testing. This is done out
156: # of curiosity at this stage as there are many other variables that you need
157: # to control with.
158: # First plots
159: plot(density(d$Female_Contributions))
160: # OK, this shows really skewed distributions
161: plot(density(d$UniqueFemaleContributors))
162: # Also this one
163: # -----
164: # Now a bit more formal test
165: ks.test(d$Female_Contributions, "pnorm")
166: ks.test(d$UniqueFemaleContributors, "pnorm")
167: # OK, this is useful. There are ties in the data. As such
168: # it would be problematic to use parametric tests. Of course
169: # the KS also shows that the distributions are not normal but maybe
170: # a significant result is normal given the amount of data.
171: # -----
172: # Since we have ties, and the assumption of normality is violated,
173: # it makes most sense to run kendal's tau for the correlation. But first,
174: # let's make a scatterplot with the data
175: plot(x = d$Female_Contributions, y = d$UniqueFemaleContributors)
176: # OK, this is so interesting. There seems to be two separate lines.
177: # This is probably due to the many zeros. Yes, the variables are problematic

```



```

178: # as, as also shown in the barplots above, many women just do not participate.
179: # OK, so let's remove the zeros
180: dtmp <- d %>% dplyr::select(Female_Contributions, UniqueFemaleContributors) %>%
181: dtmp <- dtmp[dtmp$Female_Contributions != 0 & dtmp$UniqueFemaleContributors !=
0, ]
182: plot(x = dtmp$Female_Contributions, y = dtmp$UniqueFemaleContributors)
183: # -----
184: # -----
185: #dtmp$Female_Contributions %>% table
186: #cor.test
187: #cor.test(d$Female_Contributions, d$UniqueFemaleContributors)
188: #lm(d$Female_Contributions~d$UniqueFemaleContributors) %>% summary
189: # -----
190: ```
191: # -----
192: # -----
193: ```{r}
194: # OK, here we explore two other relevant variables, namely the
UniqueFemaleContributors and
195: # FemaleParticipation
196: plot(x = d$FemaleParticipation, y = d$UniqueFemaleParticipation)
197: plot(x = d$UniqueFemaleParticipation, y = d$FemaleParticipation)
198: # OK, here there seems to be a relationship between variables but we
199: # have plenty of outliers that have an effect. I think we should do something
200: # about it.
201: # One option is to remove the outliers, especially the ones at 1. But
202: # I am a bit against that unless absolutely necessary. A common approach
203: # in dealing with outliers is to actually use robust methods. I am all
204: # for Bayesian but first let's see what we can do with the frequentist's approach.
205: corKendall(x = d$UniqueFemaleParticipation, y = d$FemaleParticipation, consistent =
TRUE)
206: corKendall(x = d$UniqueFemaleParticipation, y = d$FemaleParticipation, consistent =
FALSE)
207: # OK, both tests, show a positive correlation. Please note that we use Kendall test
208: # because we have ties in our data.
209: # -----

```

```

210: # Maybe check how the results could change if we remove the very extreme outliers
211: d$UniqueFemaleParticipation %>% describe
212: d$FemaleParticipation %>% describe
213: #
214: d$UniqueFemaleParticipation [which(d$UniqueFemaleParticipation >
mean(d$UniqueFemaleParticipation) + 3*sd(d$UniqueFemaleParticipation))]
215: # -----
216: d$FemaleParticipation [which(d$FemaleParticipation > mean(d$FemaleParticipation) +
3*sd(d$FemaleParticipation))]
217: # -----
218: d$UniqueFemaleParticipation [which(d$UniqueFemaleParticipation <
mean(d$UniqueFemaleParticipation) - 3*sd(d$UniqueFemaleParticipation))]
219: # -----
220: d$FemaleParticipation [which(d$FemaleParticipation < mean(d$FemaleParticipation) -
3*sd(d$FemaleParticipation))]
221: # -----
222: # OK, we have to remove a lot of values (for the positive limit actually). We do that for
exploratory reasons.
223: # -----
224: d1 <- d[-which(d$UniqueFemaleParticipation > (mean(d$UniqueFemaleParticipation) +
(3*sd(d$UniqueFemaleParticipation))), ]
225: d1 <- d1[-which(d1$FemaleParticipation > (mean(d1$FemaleParticipation) +
(3*sd(d1$FemaleParticipation))), ]
226: # Plot the data
227: plot(x = d1$UniqueFemaleParticipation, y = d1$FemaleParticipation)
228: # OK, the data look a bit better now. Now, let's run again the robust Kendall
229: corKendall(x = d1$UniqueFemaleParticipation, y = d1$FemaleParticipation, consistent =
TRUE)
230: corKendall(x = d1$UniqueFemaleParticipation, y = d1$FemaleParticipation, consistent =
FALSE)
231: # Again, we see a positive correlation.
232: # -----
233: # Now we are going to run a Bayesian Hypothesis Testing test. Please note we 001: use
the original variables,
234: # without removing outliers. We used the Savage Dickey approach because the non-
Savage Dickey approach
235: # had a problem with integrating with infinity.
236: # UNCOMMENT bcor <- jzs_corSD(V1 = d$UniqueFemaleParticipation, V2 =
d$FemaleParticipation, alternative = "greater")

```

```

237: # -----
238: # OK, so even the jzs_corSD did not work all of the times. When it worked it showed
stong
239: # We will investigate if we get more
240: # reliable results if we actually remove outliers
241: # -----
242: # OK, let's see what we can do with the code provided here:
243: # http://www.sumsar.net/blog/2013/08/robust-bayesian-estimation-of-correlation/
244: robust_model_string <- "
245: model {
246:   for(i in 1:n) {
247:     # We've replaced dmnorm with and dmt ...
248:     x[i,1:2] ~ dmt(mu[], prec[ , ], nu)
249:   }
250: # -----
251:   prec[1:2,1:2] <- inverse(cov[,])
252:   # -----
253:   cov[1,1] <- sigma[1] * sigma[1]
254:   cov[1,2] <- sigma[1] * sigma[2] * rho
255:   cov[2,1] <- sigma[1] * sigma[2] * rho
256:   cov[2,2] <- sigma[2] * sigma[2]
257:   # -----
258:   sigma[1] ~ dunif(0, 1000)
259:   sigma[2] ~ dunif(0, 1000)
260:   rho ~ dunif(-1, 1)
261:   mu[1] ~ dnorm(0, 0.0001)
262:   mu[2] ~ dnorm(0, 0.0001)
263:   # -----
264:   # ... and added a prior on the degree of freedom parameter nu.
265:   nu <- nuMinusOne+1
266:   nuMinusOne ~ dexp(1/29)
267:   # -----
268:   x_rand ~ dmt(mu[], prec[ , ], nu)
269: }
270: "

```

```

271: # -----
272: data_list = list(x = d[, c("UniqueFemaleParticipation", "FemaleParticipation")], n =
nrow(d))
273: # Use robust estimates of the parameters as initial values
274:   inits_list   =   list(mu   =   c(median(d$UniqueFemaleParticipation),
median(d$FemaleParticipation)), rho = cor(d$UniqueFemaleParticipation,
275:   d$FemaleParticipation, method = "spearman"), sigma   =
c(mad(d$UniqueFemaleParticipation), mad(d$FemaleParticipation)))
276: # UNCOMMENT jags_model <- jags.model(textConnection(robust_model_string), data
= data_list,
277: # UNCOMMENT inits = inits_list, n.adapt = 500, n.chains = 3, quiet = TRUE)
278: # UNCOMMENT update(jags_model, 500)
279: # UNCOMMENT mcmc_samples <- coda.samples(jags_model, c("mu", "rho", "sigma",
"nu", "x_rand"),
280: # UNCOMMENT   n.iter = 5000)
281: # -----
282: #date <- Sys.Date()
283: #save.image("mcmc.RData")
284: load("mcmc.RData")
285: par(mar = rep(2.2, 4))
286: plot(mcmc_samples)
287: summary(mcmc_samples)
288: # -----
289: # Here the results show that there is a positive correlation
290: # -----
291: ```
292: # -----
293: ```{r}
294: # The code below can be found at: https://osf.io/bg4vw/. I pasted it here because it was
easier to source like that.
295: # -----
296: # -----
297: #####
#####
298: #####
#####

```

```

299: ##### This R-code serves to compute a Bayes factor for Kendall's tau, as described in
#####
300: ##### van Doorn, J.B., Ly, A., Marsman, M. & Wagenmakers, E.-J. (in press). Bayesian
#####
301: ##### Inference for Kendall's Rank Correlation Coefficient. The American
Statistician. #####
302:
#####
#####
303: ##### To use it, input your values below for yourKendallTauValue and yourN and run
the #####
304: ##### whole script. The function call at the bottom will use your values and will compute
#####
305: ##### and print the Bayes factor in the console. The last line will plot the posterior
#####
306: ##### distribution. This analysis is also available in JASP (www.jasp-stats.org), an
#####
307: ##### open-source statistical software for Bayesian statistics with a graphical user
#####
308: ##### interface. #####
309:
#####
#####
310:
#####
#####
311: # -----
312: yourKendallTauValue <- cor(d$UniqueFemaleParticipation, d$FemaleParticipation,
method = "kendall") # Input your obtained value for kendall's tau here
313: yourN <- 100 # Input your sample size here, then run the whole script
314: # -----
315: # -----
316: # Prior specification Kendall's Tau
317: scaledBetaTau <- function(tau, alpha=1, beta=1){
318: result <- ((pi*2^(-2*alpha))/beta(alpha,alpha)) * cos((pi*tau)/2)^(2*alpha-1)
319: return(result)
320: }
321: # -----
322: priorTau <- function(tau, kappa){
323: scaledBetaTau(tau, alpha = (1/kappa), beta = (1/kappa))

```

```

324: }
325: # -----
326: priorTauPlus <- function(tau, kappa=1) {
327:   non.negative.index <- tau >=0
328:   less.than.one.index <- tau <=1
329:   value.index <- as.logical(non.negative.index*less.than.one.index)
330:   result <- tau*0
331:   result[value.index] <- 2*priorTau(tau[value.index], kappa)
332:   return(result)
333: }
334: # -----
335: priorTauMin <- function(tau, kappa=1) {
336:   negative.index <- tau <=0
337:   greater.than.min.one.index <- tau >= -1
338:   value.index <- as.logical(negative.index*greater.than.min.one.index)
339:   result <- tau*0
340:   result[value.index] <- 2*priorTau(tau[value.index], kappa)
341:   return(result)
342: }
343: # -----
344: # Posterior specification Kendall's Tau
345: postDensKendallTau <- function(delta,Tstar,n,kappa=1,var=var,test="two-sided"){
346:   if(test == "two-sided"){priorDens <- priorTau(delta,kappa)
347:   } else if(test == "positive"){priorDens <- priorTauPlus(delta,kappa)
348:   } else if(test == "negative"){priorDens <- priorTauMin(delta,kappa)}
349:   priorDens <- priorTau(delta,kappa)
350:   dens <- dnorm(Tstar,(1.5*delta*sqrt(n)),sd=sqrt(var))* priorDens
351:   return(dens)
352: }
353: posteriorTau <- function(delta,kentau,n,kappa=1,var=1,test="two-sided"){
354:   Tstar <- (kentau * ((n*(n-1))/2))/sqrt(n*(n-1)*(2*n+5)/18)
355:   var <- min(1,var)
356:   if(test == "two-sided"){lims <- c(-1,1)
357:   } else if(test == "positive"){lims <- c(0,1)
358:   } else if(test == "negative"){lims <- c(-1,0)}

```

```

359: logicalCensor <- (delta >= lims[1] & delta <= lims[2])
360: dens <- logicalCensor*postDensKendallTau(delta,Tstar,n,kappa,var,test=test)/
361:
362: integrate(function(delta){postDensKendallTau(delta,Tstar,n,kappa,var,test=test)},lims[1],lims
[2])$value
362: }
363: # -----
364: # Bayes factor computation Kendall's Tau
365: bfCorrieKernelKendallTau <- function(tau, n, kappa=1, var=1, ciValue=0.95){
366:   tempList <- list(vector())
367:   output <- list(n=n, r=tau, bf10=NA, bfPlus0=NA, bfMin0=NA)
368:   output$bf10 <- priorTau(0,kappa)/posteriorTau(0,tau,n,kappa=kappa,var=var,test="two-
sided")
369:   output$bfPlus0 <- priorTauPlus(0,kappa)/posteriorTau(0,tau,n,kappa=kappa,var=var,test="positive")
370:   output$bfMin0 <- priorTauMin(0,kappa)/posteriorTau(0,tau,n,kappa=kappa,var=var,test="negative")
371:   return(output)
372: }
373: # -----
374: # Compute credible intervals kendalls tau
375: credibleIntervalKendallTau <- function(kentau,n,kappa=1,var=1, test="two-sided",
ciValue = 0.95){
376:   nSeqs <- 1000
377:   lowCI <- (1-ciValue)/2
378:   upCI <- (1+ciValue)/2
379:   taus <- seq(-1,1,length.out = (nSeqs-1))
380:   densVals <- posteriorTau(taus, kentau, n, kappa = kappa, var = var, test = test)
381:   densVals <- cumsum((densVals[1:(nSeqs-1)]+densVals[2:nSeqs])*0.5*(taus[2]-
taus[1]))
382:   lowerCI <- taus[which(densVals>=lowCI)[1]]
383:   upperCI <- taus[which(densVals>=upCI)[1]]
384:   median <- taus[which(densVals>=0.5)[1]]
385:   return(list(lowerCI = lowerCI, median = median, upperCI = upperCI))
386: }
387: # -----
388: sampleTausA <- function(myTau,myN,nSamples = 3e3, var = 1){

```

```

389: nSeqs <- 1000
390: tauSamples <- NULL
391: taus <- seq(-1,1,length.out = nSeqs)
392: densVals <- posteriorTau(taus, myTau, myN, var = var)
393: ceiling <- max(densVals)
394: lowerB <- taus[which(round(densVals,digits=6) != 0)][1]
395: upperB <- rev(taus[which(round(densVals,digits=6) != 0)])[1]
396: # -----
397: while(length(tauSamples) < nSamples){
398:   prop <- runif(1,lowerB,upperB)
399:   propDens <- posteriorTau(prop, myTau, myN, var = var)
400:   if(propDens > runif(1,0,ceiling)){tauSamples <- c(tauSamples,prop)}
401: }
402: return(tauSamples)
403: }
404: # -----
405: bfCorrieKernelKendallTau(tau = yourKendallTauValue, n = yourN) # This function call
will carry out the computations
406: # and returns a list with the Bayes factors (regular, plussided, minsided)
407: # -----
408: # The following code plots a (simple) posterior distribution of your results.
409: layout(1)
410: plot(density(sampleTausA(myTau = yourKendallTauValue, myN = yourN),from = -1, to
= 1), las = 1, bty = "n", lwd=3,
411:   main = "Posterior Distribution for Kendall's tau", xlab = expression(tau))
412: # -----
413: # Again, it seems that there is a strong positive correlation
414: # $r
415: # [1] 0.5220499
416: # $bf10
417: # [1] 656180916077
418: # -----
419: ```
420: # -----
421: ```{r}
422: # Here, we continue with the second question.

```



```

423: # -----
424: d %>% dplyr::select(AcademicHierarchyStrict) %>% table()
425: cor.test(d$AcademicHierarchyStrict, d$ContributionsThisYear, method = "kendall")
426: d %>% dplyr::select(Workplace_SR) %>% table()
427: d$Workplace_SR %>% table()
428: # OK we see that a lot of universities have missing data -- I expected that from the variable
description but wanted to check
429: # it out. Will also investigate one more variable.
430: d$Workplace_US %>% table()
431: d$Workplace_US %>% table() %>% sum()
432: nrow(d)
433: d$Id %>% unique() %>% table() %>% sum()
434: # I actually want to make a new variable now with contributions across all years for each
unique participant
435: dplyr::filter(d, Id == "richard_dawkins") %>% dplyr::select(Year, PreviousContributions,
ContributionsThisYear) %>% dplyr::arrange(desc(Year)) %>%
436: slice(1) %>% sum()
437: # -----
438: dm <- cbind(d, ContrTotal = dplyr::select(d, PreviousContributions,
ContributionsThisYear) %>% apply(1, sum))
439: dm %>% dplyr::select(Id, AcademicHierarchyStrict) %>% unique() %>%
dplyr::select(AcademicHierarchyStrict) %>% table()
440: dm %>% dplyr::select(Id, AcademicHierarchyStrict) %>% unique() %>%
dplyr::select(AcademicHierarchyStrict) %>% nrow()
441: # -----
442: # OK, I wanted to see whether I could use the AcademicHierarchyStrict to see whether I
could use it for defining statuses but I have too
443: # many missing data. So, this is not possible.
444: # OK, but there is no way around it. Status, as a concept, can be defined in multiple ways
-- e.g., income, academic position, etc.
445: # In our case, the most **objective** way is to define status with the variable that we
have, AcademicHierarchyStrict. I mean, there are also
446: # other variables there, for example 'Job_Title_S' variable, putting status to this variable
is arbitrary. As such, I prefer to move one with my
447: # analysis using a more **objective** variable which is the 'AcademicHierarchyStrict',
than other arbitrary variable. This is because this variable
448: # has a clear meaning, and also translates in differences in status, income, etc.
449: # -----

```

```

450: dmm <- dm %>% group_by(Id) %>% summarize(ContrTotal = max(ContrTotal))
451: dmm$Ac <- NA
452: for ( i in 1:nrow(dmm)){
453:     dmm$Ac[i] <- dm[which(dm$Id == dmm$Id[i]), ] %>%
dplyr::select(AcademicHierarchyStrict) %>% slice(1) %>% unlist() %>% as.numeric()
454: }
455: # -----
456: dmmf <- dmm %>% drop_na()
457: dmmf$Ac2 <- ordered(dmmf$Ac)
458: dmmf$Ac3 <- as.factor(dmmf$Ac)
459: # -----
460: anovaBF(ContrTotal ~ Ac2, data = as.data.frame(dmmf), whichRandom = "Id", progress
= FALSE)
461: anovaBF(ContrTotal ~ Ac2, data = as.data.frame(dmmf[dmmf$ContrTotal < 101, ]),
whichRandom = "Id", progress = FALSE)
462: # -----
463: # -----
464: plot(dmmf$Ac[dmmf$ContrTotal < 101], dmmf$ContrTotal[dmmf$ContrTotal < 101])
465: points(dmmf$Ac[dmmf$ContrTotal < 101], col = 2)
466: # -----
467: d <- cbind(d, ContrTotal = dplyr::select(d, PreviousContributions, ContributionsThisYear)
%>% apply(1, sum))
468: ```
469: # -----
470: # Redo analyses
471: Some relevant websites:
472: OK, change of plans as I saw that I should report effect sizes with confidence intervals.
473: https://rdrr.io/cran/NSM3/man/kendall.ci.html
474: http://daniellakens.blogspot.nl/2014/06/calculating-confidence-intervals-for.html
475: http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r
476: # -----
477: # -----
478: ```{r}
479: d <- read.csv("edge1.1.csv")
480: # -----
481: # Prior specification Kendall's Tau

```

```

482: scaledBetaTau <- function(tau, alpha=1, beta=1){
483:   result <- ((pi*2^(-2*alpha))/beta(alpha,alpha)) * cos((pi*tau)/2)^(2*alpha-1)
484:   return(result)
485: }
486: # -----
487: priorTau <- function(tau, kappa){
488:   scaledBetaTau(tau, alpha = (1/kappa), beta = (1/kappa))
489: }
490: # -----
491: priorTauPlus <- function(tau, kappa=1) {
492:   non.negative.index <- tau >=0
493:   less.than.one.index <- tau <=1
494:   value.index <- as.logical(non.negative.index*less.than.one.index)
495:   result <- tau*0
496:   result[value.index] <- 2*priorTau(tau[value.index], kappa)
497:   return(result)
498: }
499: # -----
500: priorTauMin <- function(tau, kappa=1) {
501:   negative.index <- tau <=0
502:   greater.than.min.one.index <- tau >= -1
503:   value.index <- as.logical(negative.index*greater.than.min.one.index)
504:   result <- tau*0
505:   result[value.index] <- 2*priorTau(tau[value.index], kappa)
506:   return(result)
507: }
508: # -----
509: # Posterior specification Kendall's Tau
510: postDensKendallTau <- function(delta,Tstar,n,kappa=1,var=var,test="two-sided"){
511:   if(test == "two-sided"){priorDens <- priorTau(delta,kappa)
512:   } else if(test == "positive"){priorDens <- priorTauPlus(delta,kappa)
513:   } else if(test == "negative"){priorDens <- priorTauMin(delta,kappa)}
514:   priorDens <- priorTau(delta,kappa)
515:   dens <- dnorm(Tstar,(1.5*delta*sqrt(n)),sd=sqrt(var))* priorDens
516:   return(dens)

```

```

517: }
518: posteriorTau <- function(delta,kentau,n,kappa=1,var=1,test="two-sided"){
519:   Tstar <- (kentau * ((n*(n-1))/2))/sqrt(n*(n-1)*(2*n+5)/18)
520:   var <- min(1,var)
521:   if(test == "two-sided"){lims <- c(-1,1)}
522:   } else if(test == "positive"){lims <- c(0,1)}
523:   } else if(test == "negative"){lims <- c(-1,0)}
524:   logicalCensor <- (delta >= lims[1] & delta <= lims[2])
525:   dens <- logicalCensor*postDensKendallTau(delta,Tstar,n,kappa,var,test=test)/
526:   integrate(function(delta){postDensKendallTau(delta,Tstar,n,kappa,var,test=test)},lims[1],lims
527:   [2])$value
528: }
529: # -----
530: # Bayes factor computation Kendall's Tau
531: bfCorrieKernelKendallTau <- function(tau, n, kappa=1, var=1, ciValue=0.95){
532:   tempList <- list(vector())
533:   output <- list(n=n, r=tau, bf10=NA, bfPlus0=NA, bfMin0=NA)
534:   output$bf10 <- priorTau(0,kappa)/posteriorTau(0,tau,n,kappa=kappa,var=var,test="two-
535:   sided")
536:   output$bfPlus0 <- priorTauPlus(0,kappa)/posteriorTau(0,tau,n,kappa=kappa,var=var,test="positive")
537:   output$bfMin0 <- priorTauMin(0,kappa)/posteriorTau(0,tau,n,kappa=kappa,var=var,test="negative")
538:   return(output)
539: }
540: # -----
541: # Compute credible intervals kendalls tau
542: credibleIntervalKendallTau <- function(kentau,n,kappa=1,var=1, test="two-sided",
543:   ciValue = 0.95){
544:   nSeqs <- 1000
545:   lowCI <- (1-ciValue)/2
546:   upCI <- (1+ciValue)/2
547:   taus <- seq(-1,1,length.out = (nSeqs-1))
548:   densVals <- posteriorTau(taus, kentau, n, kappa = kappa, var = var, test = test)
549:   densVals <- cumsum((densVals[1:(nSeqs-1)]+densVals[2:nSeqs])*0.5*(taus[2]-
550:   taus[1]))

```

```

547: lowerCI <- taus[which(densVals>=lowCI)[1]]
548: upperCI <- taus[which(densVals>=upCI)[1]]
549: median <- taus[which(densVals>=0.5)[1]]
550: return(list(lowerCI = lowerCI, median = median, upperCI = upperCI))
551: }
552: # -----
553: sampleTausA <- function(myTau,myN,nSamples = 3e3, var = 1){
554:   nSeqs <- 1000
555:   tauSamples <- NULL
556:   taus <- seq(-1,1,length.out = nSeqs)
557:   densVals <- posteriorTau(taus, myTau, myN, var = var)
558:   ceiling <- max(densVals)
559:   lowerB <- taus[which(round(densVals,digits=6) != 0)][1]
560:   upperB <- rev(taus[which(round(densVals,digits=6) != 0)][1])
561:   # -----
562:   while(length(tauSamples) < nSamples){
563:     prop <- runif(1,lowerB,upperB)
564:     propDens <- posteriorTau(prop, myTau, myN, var = var)
565:     if(propDens > runif(1,0,ceiling)){tauSamples <- c(tauSamples,prop)}
566:   }
567:   return(tauSamples)
568: }
569: # -----
570:   bfCorrieKernelKendallTau(tau = cor.test(d$UniqueFemaleParticipation,
d$FemaleParticipation, method = "kendall")$estimate %>% as.numeric(), n =
length(d$AcademicHierarchyStrict))
571: # -----
572:   credibleIntervalKendallTau(cor.test(d$UniqueFemaleParticipation,
d$FemaleParticipation, method = "kendall")$estimate %>% as.numeric(),
length(d$AcademicHierarchyStrict))
573: # -----
574: cor.test(d$UniqueFemaleParticipation, d$FemaleParticipation, method = "kendall")
575: # -----
576: #' Second hypothesis
577: # -----
578: d <- cbind(d, nWords = d %>% select(Text) %>% as.data.frame() %>% apply(1,
function(x) length(unlist(strsplit(as.character(x), "\\W+")))))

```

```

579: # -----
580: d2 <- d %>% group_by(Id) %>% summarize(ContrTotal = mean(nWords))
581: d2$AcH <- NA
582: for ( i in 1:nrow(d2)){
583:   d2$AcH[i] <- d[which(d$Id == d2$Id[i]), ] %>% dplyr::select(AcademicHierarchyStrict)
584:   %>% slice(1) %>% unlist() %>% as.numeric()
585: }
586: dd2 <- d2 %>% drop_na()
587: dd2$AcH2 <- ordered(dd2$AcH)
588: # -----
589: plot(dd2$ContrTotal %>% unlist, dd2$AcH2 %>% unlist)
590: barplot(dd2$ContrTotal, names.arg = dd2$AcH2)
591: # Assumption of normality. However, we have ties so this will not work either way.
592: ks.test(dd2$ContrTotal, pnorm)
593: an <- ezANOVA(data = dd2, dv = ContrTotal, between = AcH2, wid = Id, return_aov = TRUE)
594: etasq(an$aov)
595: ci.pvaf(F = as.numeric(an$ANOVA["F"]), df.1 = as.numeric(an$ANOVA["DFn"]), df.2 = as.numeric(an$ANOVA["DFd"]), N = nrow(dd2), conf.level = 0.95)
596: #kruskal.test(ContrTotal ~ AcH2, data=dd2)
597: ```
598: # -----
599: # -----
600: # -----
601: # -----
602: # -----
603: # -----

```

REVISING A CODING ERROR IN THE DATASET

When preparing the dataset for the Boba multiverse analysis (see Supplement 11), we encountered a coding error in the original edge.org dataset for those commentators for whom we could not assign a value for the gender variable. If a thread contained a commentator for whom the gender was missing, this missing value in the gender column increased the count for the “unique female contributors” variable by +1 and it also increased the count for the “unique male contributors” variable by +1.

Consider the following example: Thread 81 contains contributions by three commentators with contributor ids: 558, 662, 664. The gender for 558 is unknown (both “Female” and “Male” are NA), and the gender for 662 & 664 is male (“Female” = 0 and “Male” = 1). In the old dataset the unique male contributors variable erroneously indicated 3 unique male contributors and the unique female contributors variable erroneously indicated 1 unique female contributor.

Thread Id	Contributor Id	Female (=1 if female, =0 if male)	Unique Female Contributors	Male (=1 if male, =0 if female)	Unique Male Contributors
81	558	NA	1	NA	3
81	662	0	1	1	3
81	664	0	1	1	3

However, a missing value in the gender variable should not have affected the count for the “unique female contributors” variable and it also should not have affected the count for the “unique male contributors.” In the revised dataset we therefore corrected the entries that were affected by this so that the example correctly indicates there are 2 unique male contributors and 0 unique female contributors in thread 81.

Thread Id	Contributor Id	Female (=1 if female, =0 if male)	Unique Female Contributors	Male (=1 if male, =0 if female)	Unique Male Contributors
81	558	NA	0	NA	2
81	662	0	0	1	2
81	664	0	0	1	2

This coding issue affected those commentators for whom no gender could be identified. It only affected those analyses in which UniqueFemaleContributors, UniqueMaleContributors or their derivatives were used in the analysis.

After we identified this coding issue in the original dataset, we reran the crowdsourced analyses twice for all analysts on the revised dataset (not just for those that could have been affected by this). Two members of the authorship team re-ran these analyses independently from one another, and both found the same results. Specifically, they found that

- for 8 analyses, neither the direction of the effect nor the significance levels have changed.
- for 1 analysis, the direction of the effect has changed from "0.000246" to "-0.001009", and the effect remains not significant either way.

- for 1 analysis, the direction of the (weak) effect has changed from "-0.0228" (not significant) to "0.3684" (now significant).

- all other analyses were unaffected by this.

A detailed comparison of the old results and the revised results can be found in the table below:

Hypothesis 1:

A#	Old result	Old p-value	Revised result	Revised p-value	Reason for revision?
1	1.059	<2e-16	1.063	<2e-16	UniqueFemaleContributors is used to filter the data
2	-1.315	0.0429	-1.315	0.0429	--
3	0.31834	0.000555	0.33311	0.000536	UniqueFemaleParticipation*10 as IV
4	0.868	<2.2e-16	0.8703	<2.2e-16	UniqueFemaleContributors as IV
5	0.5587	<2.2e-16	0.5587	<2.2e-16	--
6	-0.5872	0.0111	-0.5925	0.0103	UniqueFemaleContributors as IV
7	0.15038	<2e-16	0.15038	<2e-16	--
9	23.467	<2e-16	23.467	<2e-16	--
11	-0.0223	0.0708	-0.02331	0.0583	UniqueFemaleContributors as IV
12	9.93	< 2e-16	27.3	< 2e-16	UniqueFemaleContributors as IV
13	0.229012	< 2e-16	0.25919	<2.2e-16	UniqueFemaleContributors as IV
14	0.000246	0.946	-0.001009	0.736	UniqueFemaleContributors as IV
17	-0.0228	0.7534	0.3684156	5.544e-06	UniqueFemaleParticipation as IV
19	-0.3155	4.71e-13	-0.3155	4.71e-13	UniqueFemaleContributors as IV

Red rows indicate change in effect size direction AND significance

Yellow rows indicate change in effect size direction OR significance

Green rows indicate change in effect size magnitude. Direction and significance are unchanged.

White rows indicate no change at all in effect size, or significance levels.

Hypothesis 2:

A #	Old result	Old p-value	Revised result	Revised p-value	Reason for revision?
1	-0.1615	8.50e-05	-0.1615	8.50e-05	--
3	0.04349	0.0194	0.04349	0.0194	--
5	3.881	9.43e-14	3.881	9.43e-14	--
6	-64.38	0.52851	-64.38	0.52851	--
7	-0.22119	5.08e-05	-0.22119	5.08e-05	--
9	69.70	9.29e-09	69.70	9.29e-09	--
10	0.12150	1.36e-05	0.12150	1.36e-05	--
11	0.09032	0.1960	0.09032	0.1960	--
12	54.39	0.28738	54.39	0.28738	--
14	0.05893	0.549	0.05893	0.549	--
17	-0.05278	1.115e-05	-0.05278	1.115e-05	--
18	0.08902	0.3687	0.13252	0.1919	UniqueContributors is used to filter the data
21	0.018997	0.2421	0.018997	0.2421	--
22	-0.0374	0.582	-0.0374	0.582	--
23	-239.01	0.136	-239.01	0.136	--

Both the old dataset and the revised dataset can be accessed on the Open Science Framework here: (<https://osf.io/u9zs7/>).

Supplement 9: Qualitative analyses of explanations for analytic decisions

As described in the main text of the article, we provided a dataset to many analysts and asked them to test the two target hypotheses while carefully tracking every decision using an online platform we developed called DataExplained (Feldman, 2018; Staub 2017). By doing so, we are able to observe the roadmap of different analytical alternatives and decisions in much greater detail than ever before. In this supplement, we discuss in greater depth the steps undertaken by data analysts, and explore factors underlying the explicit and implicit decisions made throughout their data analyses.

To explore the latent factors underlying decisions, we rely on a general qualitative approach to analyze the explanations provided by different analysts. 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. To examine the exact points at which the paths diverged and forked off, we relied on DataExplained, which allows analysts to conduct their data analysis online using R while explaining at every step their decisions as they progress in the analysis (see also Feldman, 2018; Staub 2017). By asking analysts to explain their decisions and considered alternatives to the executed code, we obtain 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. We analyzed the meta-scientific data based on the way the variables were operationalized, what statistical methods were applied, and how particular variables were taken into account.

Below we review relevant literature, and describe our methodology and research design. We then report the results of the qualitative study where we seek to capture the major factors contributing to variability in analytic decisions, and outline a model describing their interplay during data analysis.

A cognitive perspective on data analysis

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 thought to be an iterative process, and intermediate output plays a key role in deciding which path to further follow. Thereby, a data analysis not only incorporates statistical or computational steps, but also cognitive processes. As Grolemund and Wickham (2014) point out, "data analyses rely on the mind's ability to learn, analyze, and understand," where each data-driven scientific work aims to "educate a reader about some aspect of reality." These analysts and readers may have different professional backgrounds and/or experiences in data analysis, as well as different mental frameworks for dealing with such tasks (e.g., forming mental models).

The concept of mental models has been studied in various research areas of cognitive science for many years (e.g., Barnes, 1944; Johnson-Laird, 1980; Norman, 1983; Seel, 2001; Weiss & Wodak, 2003). Scientists describe it as "subjective representation of the events, action, or situation a discourse is about" (Weiss & Wodak, 2003) or "qualitative mental representations which are developed by subjects on the basis of their available world knowledge aiming at solving problems or acquiring competence in a specific domain" (Seel, 2001). The process of building and interpreting such descriptions of mental models or schemas is also known as *sensemaking* (Russell et al., 1993). After being confronted with data, situated cognition and

reasoning in the sensemaking process have a considerable influence on how the data is interpreted and transformed into summary results and conclusions. Prior beliefs about a certain phenomenon may be absent, incomplete, or conflict with the apparent empirical results. Information gained from the data can help fill such gaps (if prior beliefs are incomplete), expanded (if prior beliefs are missing) or even revised (if false prior beliefs are contradicting correct information) (Chi, 2009). Hence, the data by itself can influence an analyst's beliefs, which, as a consequence, leads to different analytical choices (Paglieri 2004). A possible tool that can help researchers explore complex data and build better intuitions are appropriate visualizations (Fox & Hendler, 2011; Morton et al., 2014). Without the need for knowledge of specific programming or query languages, visual analytics might serve as efficient sensemaking tool. When being confronted with a lot of data, visualizations or visual exploration tools might help to make sense of the interplay between multiple datasets. Especially when the data is of dynamic nature (e.g., temperature profiles), appropriate visualizations can help data analysts reveal new substantial patterns, which in turn might lead to adaptations of beliefs and/or mental models (Bollier & Firestone, 2010).

That cognitive processes play a key role in data analysis has been acknowledged by some leading statisticians. Tukey and Wilk (1966) describe exploratory data analysis as the “intent to seek through a body of data for interesting relationships and information and to exhibit the results in such a way as to make them recognizable to the data analyzer.” What would be the interesting information and relationships in such case? What information is recognizable by data analyst and what will be overlooked? This is likely to be contingent on data analyst’s perception, agenda, as well as various extrinsic constraints. Moreover, at all stages of the data analysis process the outcomes of data analysis, would it be actual or potential results, have to be matched to the capabilities of people analyzing it. Thus, successful data analysis is subject to the ability to process and understand the results. Moreover, even “black box” data analysis methods like deep learning, which has gained increasing recognition in the recent years, are not useful unless the analyst can meaningfully interpret the results. Such ability relies not just within professional or technical abilities, but is part of a cognitive process inherent to the research process (Grolemund & Wickham, 2014).

Capturing the factors underlying analysis contingent results

Once it is decided which course to take throughout data analysis, it is of interest to explain the rationale behind this decision. Why should one follow this exact path, or why is this the right path to follow? Hill and Levenhagen (1995) describe this (implicit) action of communicating the perceived mental model as *sensegiving*, which eventually results in shared belief systems or consensuses (Friedkin et al., 2016). The description of the motivations underlying decisions in the context of designing a system or artefact, is referred to as Design Rationale (DR) (Lee & Lai, 1991). DR can be defined as “explanation of why an artifact is designed the way it is” and is widely discussed in the field of computer science along with many other research areas (Gruber & Russell, 1992; Schubanz, 2014). Especially in software development, it can help to effectively document and maintain artefacts (from both the UI designer's point of view as well as the technical engineer's perspective) (Guindon, 1990).

The classic concepts of a design rationale system include the existence of a design rationale database (containing design histories, reasoning, decisions, etc.). This database can be accessed with an appropriate representation schema, which elicits argumentations, decisions, or advantages and disadvantages of different options. An analyst implicitly accesses this system during the sensemaking/sensegiving processes. Conklin and Yakemovic (1991) argue that DR

can be seen as the path of decisions and selected alternatives that join the initial state (in which no decisions have been made) to the final state (in which all design decisions have been resolved). Following the metaphor of a garden with forking paths (Gelman & Loken, 2014), one could argue that any data exploration is like walking within the garden with tangled paths that might lead to different exits. In this metaphor, one could say that DR represents the full explanation as for why a certain path was preferred over others. We can describe each sub-path as a cognitive cycle a data analyst traverses, since at every one of these forks she repeatedly revisits and revises her beliefs and mental models.

Methods

Analysis platform: DataExplained

To conduct the study, we designed an online platform, DataExplained, that allows participants to run an analysis online in a RStudio environment. The platform's core consists of RStudio Server, which allows participants to conduct a data analysis using RStudio via a web browser. 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. This is essentially analyzing data in R with added transparency features. Note that since we were interested in the process via which data are analyzed, the qualitative results below include all individuals who participated in the project— regardless of whether they completed a sufficiently detailed project report and turned in code independently reviewed as free of errors (see Supplements 7 and 8).

The procedure used 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 (i.e., 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 can 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.

Each block (see Figure S9-1a and 1b) consisted of a few questions:

- Please shortly explain what you did in this block?
- What preconditions should be fulfilled to successfully execute this block?
- 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
- 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.

Edit block

✕

Please give a name to the block: *

regressions with square root and log transformation

Please shortly explain what you did in this block: *

Ran same regression as before, but with log and square root transformations of predictors.

What were the other (if any) alternatives you considered in order to achieve the results of this block?

Please describe each alternative and explain its advantages and disadvantages. By clicking on "Add another alternative", you can add additional alternatives.

Alternative

No transformation of predictors

Advantages of this alternative

Better interpretability

Disadvantages of this alternative

Potential for slightly worse diagnostic plots (heteroscedasticity, skewness of residuals)

ADD ANOTHER ALTERNATIVE

Why did you choose your option? *

I experimented with both, but will ultimately use the non-transformed data for reporting; diagnostic plots did not improve much with

What preconditions should be fulfilled to successfully execute this block? *

previous data wrangling

SHOW DIFF

DELETE BLOCK

LOAD FILES

SAVE

CANCEL

```
fit3 <- lm(comments_now_percent_change ~
log(UniqueFemaleContributors),
data = reg_dat[-244,])
summary(fit3)
plot(fit3)
fit4 <- lm(comments_now_percent_change ~
sqrt(UniqueFemaleContributors),
data = reg_dat[-244,])
summary(fit4)
plot(fit4)
```

Figure S9-1a. Example block of logs with the explanations for the code.

Edit block

Please give a name to the block:

Create different scatter plots

Please shortly explain what you did in this block:

I created a scatter plot to check the correlation between variable X and Y. In addition, I changed the color to improve the design of visualisation.

What where the other (if any) alternatives you considered in order to achieve the results of this block?

Please describe each alternative and explain its advantages and disadvantages. By clicking on "Add another alternative", you can add additional alternatives.

Alternative

Just calculating correlation coefficient Rho

Advantages of this alternative

Using statistical hypothesis testing with a p-value as output

Disadvantages of this alternative

No graphical interpretation possible, and therefore not intuitive at first sight.

Alternative

Dot-Plots

Advantages of this alternative

Good for small sets of data, as well as numerical & categorical data

Disadvantages of this alternative

Hard to construct and interpret

ADD ANOTHER ALTERNATIVE

REMOVE LAST ALTERNATIVE

Why did you choose your option?

I suspected that variable X and Y correlate because ...

What preconditions should be fulfilled to successfully execute this block?

Both, X and Y variables should be calculated based on the raw data using metric A

SHOW DIFF

DELETE BLOCK

LOAD FILES

SAVE

CANCEL

```


set.seed(170513)
n <- 200
d <- data.frame(a = rnorm(n))
d$b <- .4 * (d$a + rnorm(n))
head(d)
library(ggplot2)
ggplot(d, aes(a, b)) +
  geom_point() +
  theme_minimal()
library(ggplot2)
library(ggplot2)
ggplot(d, aes(a, b)) +
  geom_point() +
  theme_minimal()
install.packages("ggplot")
library(ggplot2)
ggplot(d, aes(a, b)) +
  geom_point() +
  theme_minimal()
ggplot(d, aes(a, b)) +
  geom_point(shape = 16, size = 5) +
  theme_minimal()
ggplot(d, aes(a, b, color = a)) +
  geom_point(shape = 16, size = 5, show.legend = FALSE) +
  theme_minimal()
d$pc <- predict(prcomp(~a+b, d))[,1]
ggplot(d, aes(a, b, color = pc)) +
  geom_point(shape = 16, size = 5, show.legend = FALSE) +
  theme_minimal()
ggplot(d, aes(a, b, color = pc)) +
  geom_point(shape = 16, size = 5, show.legend = FALSE) +
  theme_minimal() +
  scale_color_gradient(low = "#0091ff", high = "#f0650e")

```

Figure S9-1b. Example block of logs with the explanations for the code.

To help participants recall any recent changes in code, we embedded a system where it is possible to visually explore the code differences between the subsequent blocks. Additionally, participants 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.


Second, the analysts were provided with an overview of all blocks that they created during their data analysis. They could edit the blocks and reassign the respective logs to other blocks (Figure S9-2). This might be desirable if a block is not reflecting the originally anticipated goal anymore. It also allowed them to read the description of blocks following a storyline and edit the current description accordingly. At this stage, it is also possible to create new blocks that better reflect the analyst's line of thought.


Create thread-level data frame 


Adjust thread-level data frame-- add Type variable 

Adjust thread-level data frame-- deduplicate 

```
threads <- threads %>%
group_by(ThreadId) %>%
mutate(Live = ifelse(any(Live == 1), 1, Live)) %>%
ungroup %>%
distinct
threads <- threads %>%
group_by(ThreadId) %>%
mutate(Live = as.numeric(ifelse(any(Live == 1), 1, Live))) %>%
ungroup %>%
distinct
```

Adjust thread-level data frame -- follow-up from previous block, create Live_ever and live_and_not_live 

Model thread-level contributions by females with a Poisson regression 

Mean/SD for female contributions across all threads 

```
mean(threads$Female_Contributions)
sd(threads$Female_Contributions)
```


Create a data frame to look at individuals 

Figure S9-2. Fine-tuning of blocks.

Finally, in the last step of data analysis using DataExplained, analysts are asked to graphically model the workflow representing the evolution of the analysis. Initially, each analyst is presented with a straight chain of blocks, ordered by their execution. The analysts are then asked to restructure the workflow such that it better reflects the actual process. For example, iterative cycles of trying out different approaches for a sub-problem could be modeled as loops in the workflow. Figures S9-3a, 3b, and 3c show examples of workflow visualizations from analysts in the present crowdsourced project.

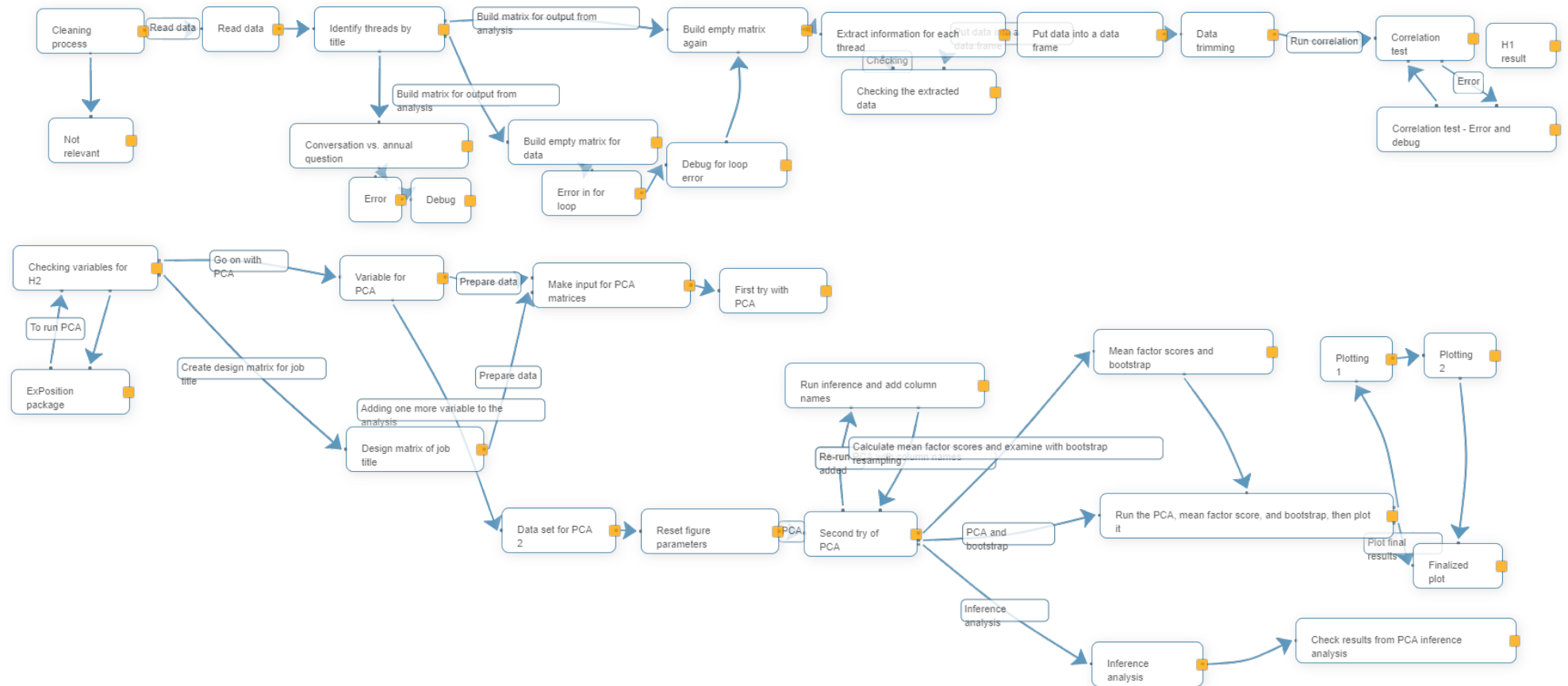


Figure S9-3a. Snippet of workflow modeled by a participating analyst.

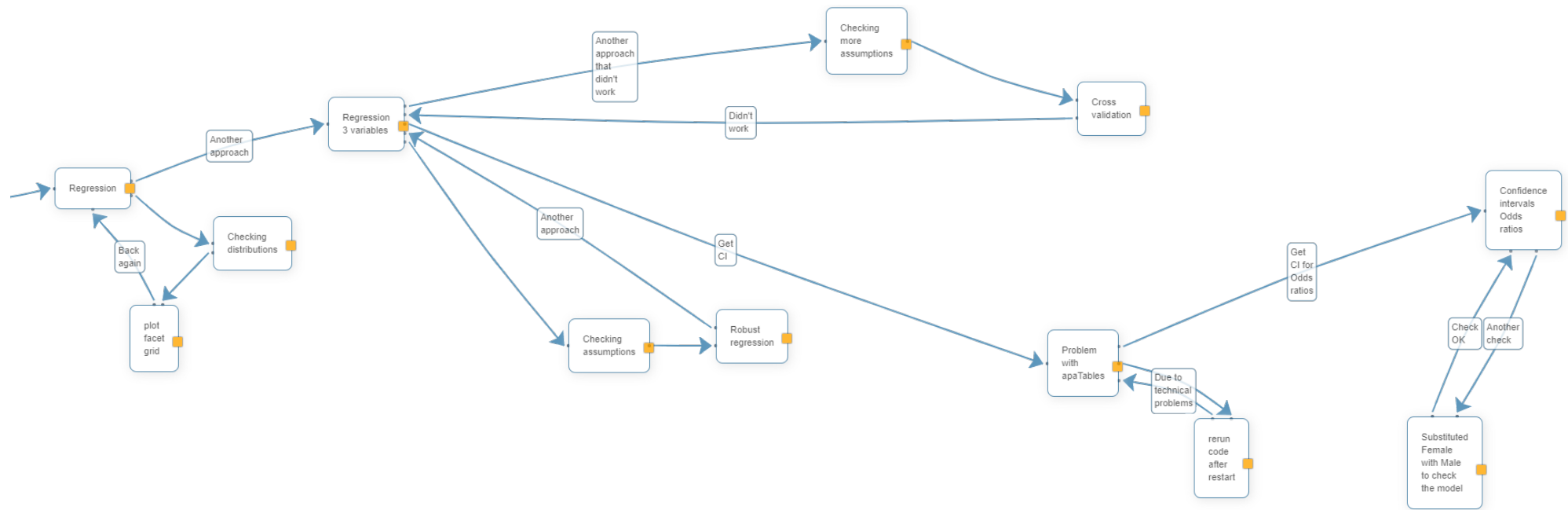


Figure S9-3b. Snippet of workflow modeled by a participating analyst.

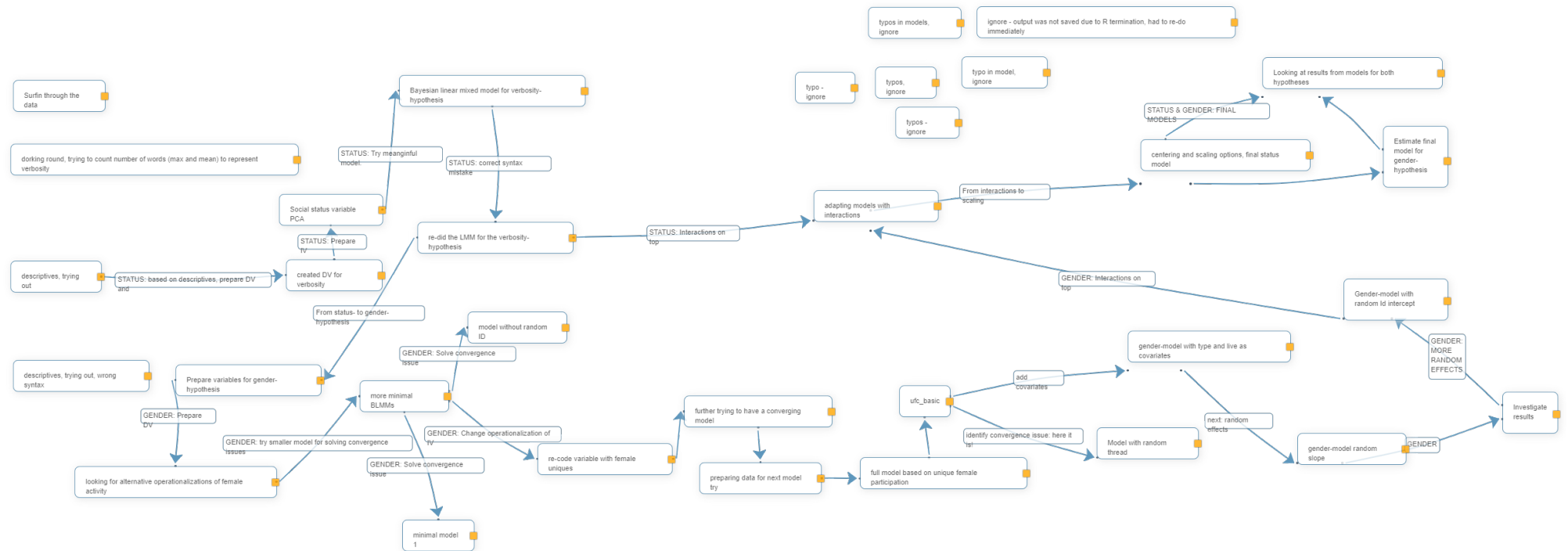


Figure S9-3c. Snippet of workflow modeled by a participating analyst.

The General Inductive approach

We utilize a qualitative research approach, which is suited for research that relies on non-structured data to describe social phenomena (Alasuutari, 2010). As described by Thomas (2006), there are four major approaches for qualitative analyses: Discourse Analysis, Grounded Theory, Phenomenology, and the General Inductive Approach.

Discourse analysis usually focuses on analyzing text as a mean of eliciting social practices and rhetoric which are emerging around topics of interest. *Phenomenology* seeks to understand the personal experiences of people who share the same experiences. The result is a coherent story describing the studied phenomenon based on the multifold of individual perspectives. The goal of a *Grounded Theory* approach is to generate a theory using a bottom-up approach based on axial coding and theoretical sampling. Last but not least, a *General Inductive Approach* seeks to develop a framework of the underlying structure of experiences or processes that are evident in the raw data. The primary goal is to allow research findings to emerge from the frequent, dominant, or significant themes inherent in raw data, without the constraints imposed by structured methodologies. This approach is more lightweight and it can lead to reliable and valid findings by following a set of standardized procedures. Even though this method is not as well-rooted as other approaches for theory building (such as Grounded Theory), it is well accepted as an approach to addressing research questions geared towards understanding underlying process.

In this study, we follow the General Inductive Approach for a number of reasons. The classical Grounded Theory approaches (Glaser & Strauss, 1967; Strauss & Corbin, 1990) are restrictive in terms of rules and procedures to follow, and often not straightforward (Partington, 2002; Thomas, 2006). This approach limits the inductive learning process to be isolated from any impact of existing theories. However, we intend to draw from the existing literature on the cognitive aspects of data analysis and the phenomenon of variability in data-analytic approaches and results. We therefore adopted a less restrictive framework for our study. The General Inductive Approach is the most suitable for this meta-scientific project, as it allows us to follow the bottom-up approach of inferring key factors, and at the same time allows us to draw on existing theories such as sensemaking.

Inductive (qualitative) coding is central to the General Inductive Approach and usually applied when there is a need to analyze volumes of verbal and written material in order to identify patterns and gain insights about the research question. The process starts with (usually) multiple researchers carefully reading the relevant materials and considering possible meanings reflected in the text. Researchers then identify text snippets that contain meaningful information and create *codes* (i.e., labels or tags) best describing the main insight of the snippet. After the researchers have refined a set of codes, they develop 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 are merged and discussed as a group. All codes as well as their memos are aggregated together into a code book. The researchers then iteratively keep refining and re-evaluating the codebook until the process resulted in a well-established and shared understanding of all the codes.

The General Inductive Approach involves five phases (Thomas, 2006). Ideally, this methodology results in the establishment of a hierarchical system of categories, where codes are low-level components and categories are high-level generalizations of the codes. Every step along these phases has certain procedures associated with them. We now describe each of them, as well as the procedures we thereby undertook throughout our analysis:

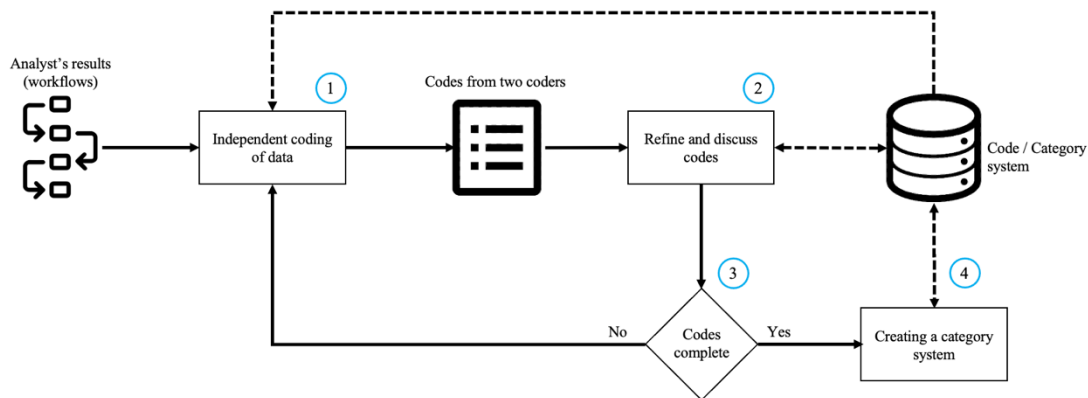


Figure S9-4. The workflow of our qualitative analysis of the quantitative analytic decisions.

Preparation of data (data cleaning)

The preliminary phase consists of transforming raw data into a common format and preparing the text for in-depth reading. In our case, we observed the procedure that data analysts followed throughout their work. Specifically, we recorded each of the commands executed and solicited their comments about the rationale of these steps. In analyses, oftentimes multiple commands address the same goal. For example, to analyze a dataset using linear regression, all variables have to be continuous. Hence, each categorical variable needs to be turned into dummy variables. The commands executed to transform all of those variables together represent one logical unit, which we call a block, with the goal of creating dummy variables in preprocessing to make the data amenable to linear regression.

To provide a useful unit of analysis, we enabled 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, sub-sequences 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. The structure of the descriptive properties originated in research on design rationale (Schubanz, 2014) and design space analysis (MacLean et al., 1991). To summarize, the main goal of a block is to provide a unit of analysis with information about its purpose, reasoning, and considered alternatives.

In our case there is no need for additional data preparation as the study is designed such that the aggregated data is already semi-structured with answers to predefined questions about the goal and the considered alternatives in blocks. Since the data from each analyst is recorded in the same format, as advised by the analysis procedure proposed by Thomas (2006), no further data cleaning is needed.

Close reading of text (coding)

This first phase consists of detailed reading of the text until the researcher is familiar with the content and gains an understanding of major themes and concepts occurring in the text. In our study, the coders sequentially went through each block of an analyst's workflow and studied the descriptive properties. Following a simultaneous coding method (Step 1 in Figure S9-4), a coder can assign multiple codes to the same attribute of text (i.e., property of the block) (Saldana, 2011). To help coders maintain consistent codes, we provided them with a searchable list of codes they had previously used. Coders could also retrieve all the explanations of the snippets annotated with the same code. This encourages a coder to continuously compare the codes and refine her reasoning. A graphical workflow for the entire sequence of blocks, refined by the analysts at the end of their analysis, provides the coders with an overview of the relationship between the blocks. Embedded in the user interface, a coder can additionally assign explanations (i.e., short memos) for every coded text segment.

Overlapping coding and uncoded text

In our analysis, one attribute can be assigned to multiple codes and much of the text may not be relevant for the research. Moreover, coders did not have strict guidelines on what codes to propose (besides the general theoretical lenses of sensemaking), since the process is inductive and therefore not restricted. This way, while analyzing text snippets, different coders could apply codes of different granularity. To help understand the context of the code, the coders could additionally explain the codes assigned to the relevant text (Fahy, 2001; Krippendorff, 2004; Kurasaki, 2000). As a result, the coded block might have different codes with a certain overlap, although the same key information was extracted by the coders.

Sometimes the answers provided by the analysts were not relevant to the research question. Hence, meaningless answers were not coded. Instead, coders were encouraged to apply codes that explain *why* the analysts provided certain answers. We asked the coders to apply codes to block attributes in order to ease the task of interpretation. For example, the coders were asked to elicit the goal of the block, the considered alternatives, or why a certain alternative was preferred. Additionally, to capture the general purpose of a block, coders could also assign codes describing the general goal of the block.

Creation of categories

In this phase coders collaboratively defined codes and discussed categories by summarizing and aggregating codes by their meaning (Step 2 in Figure S9-4). This was reached through a discussion where the meaning of the codes was clarified and the semantically identical codes were merged. Further, based on the list of codes, the coders collaboratively constructed a category system – a high level organizing abstraction which summarizes the codes (Step 4 in Figure S9-4). Each refined category was provided with a memo, which summarized the coders thoughts and/or possible relations to other codes/categories. These memos not only served as justifications for the category, but also facilitated future revision and refinement of the category system. As Creswell (2002) suggests, this newly emerged list of categories should serve as new organizing scheme for

coding. This is instrumental in inferring the categories based on the codes after a further iteration. In our case, the (updated) list of categories served as new coding scheme for coding in the subsequent iteration. This scheme is then to be applied to another subsample of the data, where coders can draw on the reasoning from memos when applying codes. Nevertheless, they can still come up with new codes, which are then either assigned to an existing category or begin the foundation for a new category.

Continuing revision and refinement of category system

Each iteration of coding ended with revision and refinement of the category system (Step 2 in Figure S9-4). The number of total assigned codes to a category as well as their prevalence could indicate the importance of categories. A category with only few assigned codes might indicate that this category is not well grounded. In such cases, we considered merging this category with a more robust category (i.e., with more frequently occurring codes assigned to it) or removing the infrequent category. Miles et al. (2014) described the process of grouping initial codes into a smaller number of categories as *pattern coding*. We eventually reached the point where no new codes and categories emerged from the new subsample of data (Step 3 in Figure S9-4). At this point, if the coders believe that each element accounting for the inherent variation in data analysis is captured by a category, the iterative coding is finished. A high percentage of agreement among the coders (i.e., proportional agreement) guaranteed not only a common understanding of the coding scheme, but also showed a level of high agreement when applying them.

Establishing reliability of the qualitative codings

There are many ways to evaluate the reliability for models developed in qualitative analysis. The literature offers several ways to evaluate intercoder reliability or intercoder agreement (Campbell et al., 2013). According to Campbell et al., the use of such statistics for qualitative analyses aiming for systematic and rule-guided classification and retrieval of text are less imperative. As a consequence, simple proportion agreement (percentage of agreement among coders) is seen as a reasonable approach (Kurasaki, 2000). Moreover, some researchers suggest that looser standards are permissible in exploratory studies (e.g., Hruschka et al., 2004; Krippendorff, 2004). In order to guarantee high reliability of the emerged final categories in this study, we applied both qualitative and quantitative measures of reliability:

Independent parallel coding. Two coders independently developed a set of codes (Step 1 in Figure S9-4). These two sets were compared and merged into a combined set (Step 2 in Figure S9-4). When the overlap between the codes was low, the coders discussed and clarified each code in order to reach a more robust set of codes. This procedure also resembles the negotiated agreement approach proposed by Campbell et al. (2013).

Check on the clarity of categories. Two additional independent coders (previously not involved in coding) were introduced to the set of codes supplemented with explanations and examples. All coders were then asked to code a new subsample of data using the system of codes. Thereafter, if they came up with new codes the code book was refined and translated into a new coding scheme. This coding scheme is then used for coding new data in another iterative cycle.

Calculation of interrater agreement. To measure the agreement among coders, we calculated the proportional agreement and Cohen's Kappa after each iteration (i.e., in Step 2 in Figure S9-4).

Our qualitative approach was guided by the overall objective of the project, thus focusing on the question of what factors are contributing to the variability in data analyses. By doing so, we did not explicitly rely on any theory but let the findings arise directly from the interpretation of the raw data. We interpreted the blocks using all available information, such as the workflow of blocks and their descriptions (taking into account future and past blocks), as well as analysts' comments and source code. Thus, our "coding filters" were broadly split into two areas. First, the objective output of a block, such as the method selection, the revision of code, or the task constraint. These factors are objective and all analysts face them equally throughout data analysis. Second is the subjective decision making process involved in data analysis. Factors such as personal beliefs, experiences, knowledge, or intermediate insights which inform the next steps in a data analysis and differ among analysts. Due to the developed construct of "blocks" as well as the graphical representation of workflows, we had the necessary information to explain the rationale for every step in a data analysis.

First, two coders coded the blocks in a sequential manner, proceeding through blocks in their chronological order (Figure S9-5). For every applied code they provided an explanation for the code. As a result, every block was annotated with i) codes, ii) possible explanations, and iii) a reference to a relevant snippet, were it the analyst's verbal explanation or an executed command.

After both coders finished coding the blocks from a predefined subsample, the codes were grouped together. Next, the coders collaboratively refined and discussed each of them. As a result, similar codes were merged together, whereas overly general codes were split into more fine-grained codes. For each code, the coders created a short explanation in the form of a memo and provided some examples where this code has been applied. The resulting code book (codes along with memo and examples) was then used for the subsequent coding iteration.

When the inter-rater agreement is high enough (proportional agreement and Cohen's kappa $> .70$), the code book was presented to two additional coders. After they learned and refined the codes together with two initial coders, all four coders coded a new subsample and verified that the codes are suitable to describe the rationales of the data analysts. In this phase, the code book is further refined and new subsamples are coded until the agreement among all four coders is high enough. In order to proceed to the next step, all coders iterated four times until the proportional agreement among them reached at least 50%.

Code block

Please give a name to the block [title]

plotting and regression with subset

Please shortly explain **what** you did in this block [goal]

Plotted percent change in comments against number of unique female contributors. Tried a regression without the major outliers.

What where the other (if any) alternatives you considered in order to achieve the results of this block?

Please describe each alternative and explain its advantages and disadvantages

Alternative [alt1]

Plot only linear relationships

Advantages of this alternative [adv1]

More strictly represents regression.

Disadvantages of this alternative [dis1]

doesn't help to see non-linear trends.

Why did you choose your option? [reason]

wanted to explore data a little more

What preconditions should be fulfilled to successfully execute this block? [prec]

data wrangling

```
> ggplot(reg_dat, aes(x = UniqueFemaleContributors)) + geom_freqpoly()
> ggplot(reg_dat, aes(x = UniqueFemaleContributors)) + geom_freqpoly(bins = 15)
> ggplot(reg_dat, aes(x = UniqueFemaleContributors)) + geom_freqpoly(bins = 10)
> ggplot(reg_dat, aes(x = UniqueFemaleContributors)) + geom_freqpoly(bins = 8)
> ggplot(reg_dat[244, ], aes(x = UniqueFemaleContributors,
> y = comments_now_percent_change)) +
> geom_smooth(method = "loess") +
> geom_point(alpha = 0.5)
> subset <- reg_dat[244, ] %>%
> filter(comments_now_percent_change < 500)
> ggplot(subset, aes(x = UniqueFemaleContributors,
> y = comments_now_percent_change)) +
> geom_smooth(method = "loess") +
> geom_point(alpha = 0.3)
> fit5 <- lm(comments_now_percent_change ~ UniqueFemaleContributors, data =
> subset)
> summary(fit5)
> plot(fit5)
```

Coding

exploratory data analysis x

reason

wanted to explore data a little more

perceived suitability of the method x pers

adv1

personal interest

personal preferences

exploratory data analysis x outliers x Add code

goal

Tried a regression without the major outliers.

ADD ANOTHER CODE

SHOW DIFF CANCEL SAVE

Figure S9-5. Coding interface for a block.

Results

Two coders followed three coding cycles in order to build a sustainable coding scheme. After each iteration, they discussed any discrepancies in their results and refined the codes. In the first iteration, both coders independently coded the 275 blocks of ten different analysts, to come up with an inclusive list of initial codes. As a result, they constructed 88 codes describing various factors contributing to variability in data analysis. After eliminating duplicates (i.e., semantic synonyms) and insufficiently justified codes, they were left with 30 codes. To check whether these codes were inclusive and complete, an additional subsample of 49 blocks corresponding of five analysts was then analyzed. During this iteration, coders realized that some codes were too general and needed further refinement (i.e., either split the code into more detailed codes or delete the code entirely, as other codes may already substitute it). Therefore, they reviewed the blocks where rather high-level codes were applied and refined them to be more precise.

The coding scheme for the last and third iteration consisted of 31 codes and 41 blocks. The coders then coded another subsample (21 blocks), however the codebook remained unchanged (i.e., they neither came up with new codes nor deleted any of already existing codes). The proportional agreement of the two coders after the last iteration was 72%, with a Cohen's Kappa of .70. The resulting codebook was then presented to two new coders. They were provided with code-memos and examples of when (and when not) to apply each of the codes, and clarifications of any differences between related codes. All four coders then discussed the codes and clarified them with their corresponding memos. Following this, the coders independently coded another subsample of 22 blocks. All four coders then discussed the results of their coding and updated the codebook

accordingly. The coders then coded another subsample of 9 blocks. After the third iteration performed by all four coders, the percentage agreement reached 52.6%, and the codebook was finalized. Since there were no more disagreements at this point, there was no need for an additional coding iteration. At the end, the final codebook consisted of 31 codes. The four coders collaboratively grouped the codes together and created a category system with ten categories (see below).

Codebook

In the following we describe the categories and list the corresponding codes. We provide a succinct explanation of the codes, a few examples of participants' corresponding comments, and a short discussion of how the categories contribute to the variability in data analysis. Some actions conducted by data analysts are not a result of one isolated consideration but rather a blend of multiple factors involved in their decision making. Therefore, we often attributed multiple codes to a given analyst comment. In the examples provided here though, we discuss each comment in the context of the single most descriptive code.

Category	Codes	Category Description
Data	<ul style="list-style-type: none"> • Data constraint: Any constraint imposed by the nature of data • Data quality: Any objective metrics of data quality such as completeness, bias, distribution etc. • Feature engineering: Adding new features (aka variables/columns /attributes etc.) which are a function of existing data. • Preprocessing: Any steps performed to preprocess the data (e.g., installing packages/libraries, removing outliers, organize data, etc.) 	<p>This category reflects all activities and considerations related to data. Data might have objective constraints such as format, missing values, or size. Also, data transformation (i.e., feature engineering) and data preprocessing are data related activities which are not only changing the data, but might channel data analysis in certain direction.</p>
Examples of participant comments:		
<ol style="list-style-type: none"> 1. “There's no variance in number of comments made. Data also has a temporal structure [so] that last analysis ignored” (data constraint) 2. “I don't think there is enough data to parameterize this model” (data constraint) 3. “Created paired changes in status normalized by the changes in the same period among people who did not change status” (feature engineering) 4. “Go through each row in original data, and only extract the first conversation of each thread” (preprocessing) 		

Contribution to variability in approaches:

Data constraints limit and channel data analysis into certain direction. While sometimes these constraints cannot be ignored (e.g., missing data, data size), it is a matter of expertise and experience to notice the problem in other cases. For example in (1), the analyst realized that the data is temporal. This made the results he/she had obtained invalid and resulted in a different approach being adapted instead. Another example is subjective decisions, such as what data to select as a subset of (4). In this case, the conclusions were derived based on this data. If the analyst were to sample the data differently, and pick random conversations in each thread, the results could be different. Furthermore, analysts often transformed variables to be able to operate with more informative features (aka feature engineering). As it can be seen from (3), the way variables are transformed is a function of the analyst's internal hypothesis about the best way to operationalize the problem and may impact the subsequent analyses.

<p>Task</p>	<ul style="list-style-type: none"> • Task constraint: Task constraint is related to the limitations imposed by the task the analyst is performing (requirement for the task). For example, if the task is to report on certain measures or to produce a result up to certain deadline. • Complexity constraint: Complexity constraint represents cases where the analyst considers the complexity of alternatives or performed methods. A method might be objectively better but still avoided due to the analyst's reluctance to engage in complicated data analysis process. This code is related to "effort constraint". However while the "complexity constraint" is related to the perceived complexity of the method (i.e., how complicated is it to execute), the effort constraint is related to the effort associated with the alternative, which does not necessarily results from the complexity of 	<p>Task constraint is related to the task which has to be accomplished during data analysis. This task could be either answering a hypothesis, or an exploratory analysis aiming to produce potential research questions that could be answered with the data at hand.</p> <p>Complexity constraint represents cases where an analyst is considering the complexity of alternatives or performed methods. A method might be objectively better but still avoided due to the analyst's reluctance to engage in complicated data analysis processes. On the other hand, task constraint is related to the limitations imposed by the task the analyst is performing (i.e., task requirement). For example, when the task is to report on certain measures or to produce a result up to a certain deadline.</p>
--------------------	--	---

	the method. Another relevant code is a "methodological constraint." This code relates to the objective constraints imposed by the requirements of a method.	
Examples of participant comments:		
<ol style="list-style-type: none"> 1. “not within scope of hypothesis” (task constraint) 2. “That the project requires the reporting of effect sizes and my approach - based on Bayes factors - cannot do that” (task constraint) 3. “complicated getting data into tm (date) format” (complexity constraint) 4. “More difficult to keep track of things” (complexity constraint) 		
Contribution to the variability in approaches:		
<p>When an analyst considers various alternatives for analyzing the data, task constraints and goals play a key role. For instance, if the task requires to report certain measures (2), or if the considered method requires the data to be in a certain format (3), the analyst will prefer certain analytical alternatives. Moreover, analysts might not proceed with exploring some ad-hoc hypotheses that arose during analysis if they seem to be not within the scope of the task (1). Nevertheless, some of them could be helpful for answering the core questions of the overall analysis.</p>		
Problem perception	<ul style="list-style-type: none"> • Uncertainty about the problem: In this context by problem we mean the phenomenon which is under investigation. A problem the analyst studies might be ambiguous in its nature for different reasons. In addition, any uncertainties expressed with regards to the problem setting (e.g., if an analyst is not sure what is the meaning of a variable in dataset, how the 	<p>This category refers to the problem the analyst is studying. This problem could be a hypothesis under investigation or an exploratory analysis. The perceived understanding of “problem mechanics” impacts an analyst’s actions and informs intermediate steps throughout the data analysis. A problem an analyst studies might be ambiguous in its nature for different reasons, such as loose specifications or different interpretations of certain aspects. In addition, any uncertainties expressed with regards to the problem setting (e.g., if an</p>

	<p>data was collected, or how to interpret the results)</p> <ul style="list-style-type: none"> Perceived understanding of the problem: This code is applied when analyst is following a procedure due to the perceived logic of the problem. This code is mostly applied when a justification for the action is given with regards to the problem. Note, this code is different from the perceived understanding of reality. While perceived understanding of reality is reflecting a general context, understanding of the problem reflects a concrete problem the analyst currently deals with and the sensemaking process that occurs. Selecting features/variables belongs to this code. Intuition about the problem: Intuition is a "gut feeling" that results out of prior knowledge or by inference from personal experiences, feelings and preferences. Intuition in this case refers to intuitions about future actions. 	<p>analyst is not sure what the meaning of variable in dataset is, how the data was collected, or how to interpret the results) might affect the data analysis. Moreover, analysts often have an intuition about a problem. This kind of a "gut feeling" results from the prior knowledge or by inference from personal experiences, feelings and preferences. In data analysis, intuition might come into play when the analyst automatically relies on it, in order to inform next intermediate analytical steps.</p>
--	---	---

Examples of participant comments:		
<p>1. “The scale is ordinal, but it's unclear to me how different each level is from the other - how much different is an experienced graduate student from a post-doc? An associate professor vs. a full professor? It seemed better to simply recognize them as nominal categories” (uncertainty about the problem)</p> <p>2. “It's hard to separate being female from many other factors that may also be the result of being female. Wanted to focus on a clean overall result without many controls. As noted in one alternative, couldn't come up with a reliable way to know if a female participant knew if there were other females in the conversation except for authors. There wasn't enough variation in number of times participating to use that to define active participation” (uncertainty about the problem)</p> <p>3. “If hypothesis is that seeing women talk draws other women to be more active, the woman posting can only see that in regular discussions, not in annual conversations” (perceived understanding of the problem)</p> <p>4. “These two variables atm seemed to be a good choice for the verbosity-operationalization, after going through all the language-variables created from the liwc” (intuition about the problem)</p>		
Contribution to the variability in approaches:		
<p>Research questions often hypothesize about high level constructs. Operationalization of these constructs is not always clear (1-2). This is where the analyst is mostly relying on her or his intuition about the problem. For example in (4), the analyst is stating that intuitively there are two variables in the dataset that might be a good representation of the construct of verbosity. Differently from intuition about a problem that can be seen in (4), in (3), the analyst expresses her perceived understanding about the problem. This means that there is much more certainty about understanding of the mechanics of the problem domain. For example in (3), there is a clear statement that the researched phenomenon cannot be observed in certain data.</p>		
Knowledge	<ul style="list-style-type: none"> • Perceived course of action: The analyst performs an action in order to be able to continue the way she intends (e.g., when analyst states a clear path to operationalize the problem - "Do A in order to do B"). • Personal knowledge: Analyst's knowledge or prior experiences in performing an action she does (e.g., refers to past analyses, claims to be 	<p>The knowledge and experiences the analyst possess (e.g., when she refers to past analyses, claims to be familiar with a concept, or consequences of possible actions). The code “perceived course of action” describes a situation where the analyst performs a certain step in order to be able to further follow in a certain direction during the analysis. For example, when the data is transformed into a certain format in</p>

	<p>familiar with a concept, or consequences of possible actions)</p> <ul style="list-style-type: none"> • Method preference: Analyst's preference for certain methods. This can be either due to professional background/education or commonly faced problems. For example Bayesian statisticians prefer certain methods while some other researchers' favor frequentist methods. • Expertise: Decisions or actions that reflect professional knowledge and experience. For example, when analyst is considering that while applying a certain method, one has to be careful of certain aspects such as assumptions or limitations. • Effort constraint: Effort constraint represents cases where the anticipated effort prevents the analyst from taking certain actions/decisions during data analysis. This can be either due to time/complexity constraint or because the perceived benefit versus invested effort do not make it attractive ("too much work to be done"). 	<p>order to be able to apply an intended method (e.g., binarization of the outcome variable in order to perform a logistic regression). Furthermore, we observed expertise through decisions or actions that reflect professional knowledge and experience. For example, when an analyst is considering that when applying a certain method, one has to be careful of certain aspects such as assumptions or limitations. Awareness of the assumptions as well as consideration of methodological alternatives and their limitations, were seen as an indication of expertise. Last, effort constraint represents cases where effort prevents an analyst from taking certain actions/decisions during data analysis. This can be either due to time/complexity constraint or because the perceived benefit versus invested effort does not make it attractive (or "too much work to be done" as it was often reported).</p>
--	--	---

Examples of participant comments:

1. “Took data where each observation was a participant, and summarized it down to a dataset where each observation is a conversation. I wanted to be able to study things at the conversation level” (**perceived course of action**)
2. ”these packages have been useful in my past analyses” (**personal knowledge**)
3. “Tried to run a Bayesian Hypothesis Test using the functions in BayesMed but it did not work” (**method preference/methodological constraint**)
4. “Models need to converge, and the choice of model terms cannot be data-driven since that would render the p-value for the χ^2 test meaningless due to the garden of forking paths” (**expertise**)
5. “More columns, harder to do tests based on blocks of variables” (**effort constraint**)

Contribution to the variability in approaches:

In (1) the analyst is summarizing the data to the conversation level in order to conduct further analyses on this level. Since the aggregation might often lead to information loss and lead the whole analysis in a certain direction, the intended course of action contributes to the variability in data analysis. The same is true for the analyst's personal knowledge (2), method preference (3) and expertise (4), which all play a key role in predefining the course of data analysis. Lastly, the effort constraint is the factor that often undermines the depth of analysis. Like in (5), analysts often choose to avoid certain activities because they are time and effort intensive and will require too much of his or her resources.

Belief	<ul style="list-style-type: none"> • Perceived understanding of reality: The perceived understanding of the reality is a complementary factor to beliefs and interests. Data analysts may have an implicit cognitive schema about “how things work” in the real world. This understanding is not directly about the problem which is under investigation but rather about a general state of the world. • Personal assumption: Any personal assumptions the analyst makes. For example, the analyst dropped most of the PhDs from his or her analysis as they were not expected to influence the final result much. • Personal interest: Actions driven by personal interest of the analyst (e.g., curiosity, 	<p>This category describes the tacit belief system of the analyst. Any personal assumption the analyst makes or action driven by personal interest of the analyst (e.g., curiosity or choices which relate to personal perceived rationales) might be categorized as part of the belief system. It is different from explicit knowledge by being tacit by nature. It might be the personal belief (agenda) for an analyst to prove that a certain hypothesis is correct (e.g., the presence or absence of bias against women in scientific discussions). Analysts may have preferences or intentions to perform an action the way they think is best for them. These can be driven by various personal factors. Such predispositions might play a key role in the way the data analysis is conducted even though no explicit traces can be observed in the data analysis results. The perceived understanding of the reality is a complementary factor to beliefs and interests. Data analysts may have an implicit cognitive schema about “how things work” in the real world. This understanding is not directly about the problem which is under investigation but rather about a state in the grand scheme of things.</p>
---------------	---	---

	<p>choices which relate to personal rationales)</p> <ul style="list-style-type: none">• Personal preferences: Analysts may have preferences or intentions to perform an action the way they think is best for them. These can be driven by various personal factors. If the preference is for a (statistical) method, we apply only the code "method preference."	
--	--	--

Examples of participant comments:

1. "I chose this option because there was no way to determine the value of job titles, however I think they are important. A director or a president has higher status than a graduate researcher and this should be reflected in the status" (**perceived understanding of reality**)
2. "Because the hypothesis is based on verbosity of users and not individual posts. My option assumes that total characters of each user is a strong metric for their overall verbosity" (**personal assumption**)
3. "interested in seeing how different disciplines have different gender breakdowns" (**personal interest**)
4. "Habits: I mostly start data analysis with such first steps" (**personal preferences**)
5. "I believe it's more robust" (**belief**)

Contribution to the variability in approaches:

Perceived understanding of reality describes the mental models of an analyst. For example in (1), once the analyst encountered an uncertainty, she relied on the perceived understanding of the importance of job titles. Hence, this variable was transformed into ordinal and included in the model. Other analysts would overlook this variable, and most likely even - if not - operationalize it differently (e.g., interpret the hierarchy of job titles otherwise). The analyst in (2) also makes a personal assumption while deciding to operationalise verbosity through total number of characters. Additionally, when conducting an analysis, scientists are sometimes drifting from the core hypotheses in order to answer questions which are of their own interest, as exemplified in (3). The insights gained from this exploration may inform the main analysis and have impact on the results. Moreover, personal preferences and beliefs (4) inform the analysis and lead it in certain directions. For example, if one analyst starts her data analysis with data visualisation and exploration, the insights gained during this step might divert her from the initially anticipated course of analysis.

Exploratory data analysis	<ul style="list-style-type: none"> • Exploratory: Any exploratory steps performed by the analyst. This is related to exploratory data analysis and can describe activities focused on data or model exploration. • Visualisation: Any kind of graphical visualisation / plot the analyst does. This is often related to the code "insight generation" or "exploratory analysis" 	<p>Exploring and understanding the data. This is related to exploratory data analysis and can describe activities focused on data or model exploration. For instance, data plotting and visualisation is often part of the exploratory data analysis where an analyst is attempting to understand data properties and their behaviour. This is also often related to the code "insight realization," since visualization often leads to new insights throughout the data analysis. Exploratory data analysis is well acknowledged as a cornerstone in data analysis (Tukey, 1977) and considered as a highly interpretative component that may influence the direction of further analyses.</p>
Examples of participant comments:		
<ol style="list-style-type: none"> 1. "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" (exploratory) 2. "Selected status metrics iteratively: identified several, plotted them, removed redundancies, plotted again and checked for correlations" (exploratory) 3. "Looked at the univariate distributions for each column (Hmisc::describe()). Plotted the number of conversations/year. Plotted distribution of female participation as a density plot, then created scatter plots looking at male vs. female contributions; and # female contributors vs. female participation" (visualisation) 		

Contribution to the variability in approaches:		
Exploratory analysis is very common and occurring in many stages of data analysis. Even when the analysis is confirmatory by nature, analysts very rarely follow a predefined path to analyse the data. Usually there is continuous exploration of the data that has impact on the way the analysis is conducted (aka adaptive data analysis) such as in (1-2). Visualisation (3) is one of the most powerful tools to explore data and is widely used in data analysis.		
Confirmatory data analysis	<ul style="list-style-type: none"> • Revision of findings: Revision of findings due to <i>new</i> insights or ideas. Often related to the code "insight realization" • Confirmatory measure: Analyst tend to confirm their (intermediate) results in different phases of their analysis. 	Reassures that the output makes sense and is correct. For example, that the data is indeed distributed according to an assumption, the results are within the expected range of values, or that the results are credible. Another example is the revision of findings due to a new insight or idea. Often the analyst has an insight or hypothesis about the problem and seeks to reconfirm it by checking whether the data corresponds.
Examples of participant comments:		
<ol style="list-style-type: none"> 1. "Ran the code from beginning to the end again, looked at the plots and rethought the modeling" (revision of findings) 2. "Re-ran the code to double check whether things are ok and to look in detail at the effect sizes and estimates" (revision of findings) 3. "Did a check with another analysis where I substituted Female with Male to reassure that the reversed coding of that variable didn't affect the R2" (confirmatory measure) 4. "Checked that the number of observations in the women-only subset was in line with what was expected" (confirmatory measure) 		

Contribution to the variability in approaches:		
<p>The category refers to reflections on the reached data analysis results. As stated by (1,3), often revision of the model sparks new insights and leads to remodeling steps. Experienced analysts often examine the intermediate results in order to assure that the outcomes are not flawed and make sense. This sensemaking process often leads to updating in the perceived understanding of the problem and causes analysts to reconsider the course of analysis.</p>		
Methodology	<ul style="list-style-type: none"> • Uncertainty about the method: If analyst is not sure whether the employed method is the correct one for her objectives or another method would fit better • Methodological constraint: A methodological constraint related to the limitations imposed by considered methods or approaches. For example, assumptions of normality or homoscedasticity have to be fulfilled in order to apply certain methods. • Interpretability constraint: Analysts have a subjective judgement for the interpretability of methods or approaches. This is a subjective constraint 	<p>Describes the methodological aspects of the conducted analysis. The methodological decisions might range from high level methodology to be used (e.g., Bayesian vs. frequentist statistics) up to concrete decisions, such as how to operationalize the variables. Furthermore, analysts sometimes are not sure whether the selected method is the correct one for their objectives. Whenever we found evidence for such uncertainty, we related this to “methodology.” Lastly, a “methodological constraint” is related to the limitations imposed by considered methods or approaches. For example, the assumptions of normality or homoscedasticity have to be fulfilled in order to apply certain methods. Analysts have a subjective judgement for the interpretability of methods or approaches. This is a subjective constraint.</p>

Examples of participant comments:

1. “A mix of harder to model and not sure about the right assumptions” (**uncertainty about the method**)
2. “Unsure about whether I missed a covariate in the model and whether I need to change to a model accounting for the fact that the hierarchy variable is ordinal” (**uncertainty about the method**)
3. “Variables need to be at least ordinal”, or, “model doesn't converge” (**methodological constraint**)
4. “This [method] seems simple, common-sense, and easy to interpret” (**interpretability constraint**)

Contribution to the variability in approaches:		
<p>A decision of what method to apply is important and is often influenced by considerations, such as method sensitivity, robustness to assumption violations, and underlying approaches (e.g., frequentist vs Bayesian). Additionally, when the method is hard to interpret (4) or mathematical modeling such that an alternative method could be applied is challenging (1-2), an analyst often opts for simpler model. Hence, the uncertainty about alternative methods often results in analysts reusing the same, more familiar method across different datasets, even when they are aware of potentially more suitable methods. Since the statistical assumptions of methods are often open for discussion, analysts are often not sure how restrictive they should be with regards to this.</p>		
Insights	<ul style="list-style-type: none"> • Insight realization: This code describes a situation where the analyst generates new insights, hypotheses, or ideas, due to the applied method/approach or during the data analysis more generally. This code can be seen as an evidence of sensemaking. • Action driven by insight: Analyst's personal insights may drive certain actions to be followed (e.g., running a correlation test on two variables of interest emerged from the generation of an insight). Often related with the code "Insight realization" 	<p>Reflects the insights gained throughout the data analysis. Insight realisation is a code that describes a situation where the analyst generates new insights, instant hypotheses or ideas, due to the applied method, approach or throughout the data analysis in general. This code can be seen as an evidence of sensemaking. Analysts' personal insights may drive certain actions to be followed (e.g., run correlation test on two variables of interest that emerged from the insight generation).</p>

Examples of participant comments:

1. “I compared Threads to Job Title along with PhD Ranking, and found as prestige of Job Title increases, number of Threads increases, and this is especially true for higher PhD Ranks” (**insight realization**)
2. “Turned entries into paired data for people with word count and status. Thing repeated the process because I checked and realised sometimes people had more than one answer to an annual question” (**action driven by insight**)
3. “Prepared the individual entries for testing H2 based on the realisation that WC (*i.e.*, *word count*) is sensitive to what year the communication was in” (**action driven by insight**)

Contribution to the variability in approaches:		
<p>One of the reasons for a data analysis to develop in a certain direction are intermediate insight realizations analysts have in the process. For example, (2) had an insight, that, as prestige of Job Title increases, number of Threads increases. Such realizations inform the decisions this analyst makes throughout her data analysis. For example the insight (3) had about “WC (<i>i.e.</i>, <i>word count</i>) is sensitive to what year,” triggered the restructuring of data, in order to better account for this phenomenon.</p>		
Coding skills	<ul style="list-style-type: none"> • Code quality: Actions performed to enhance the objective quality of code (e.g., reorganize, refactor, comment, etc.) • Debugging: Code executed for debugging / corrective measures. 	<p>Since we explore a case where the data analysis is conducted without a user interface mediation but through R coding, these actions reflect coding skills of the analyst. Code quality relates to the measures undertaken by an analyst to enhance the objective quality of code (e.g., reorganize, refactor, comment, etc.). Debugging code relates to activities whose purpose is to find an error in code that presents unexpected or seemingly incorrect results. It also includes activities related to testing whether corrections were effective.</p>
Examples of participants comments:		
<ol style="list-style-type: none"> 1. “It's cleaner code since I only use it for a few variables” (code quality) 2. “Rewrote and commented the code (final pretty version.R) so that it was better for sharing, then reran the analysis of the code” (code quality) 3. “Caught an error, rerunning analysis with error fixed” (debugging) 4. “Troubleshooting the aggregation by participant” (debugging) 		

<p>Contribution to the variability in approaches:</p>
--

<p>The major contribution of code quality to the variability in data analysis is through the complexity that it introduces. On one hand, analysts who write less clean and not well documented code may also not cross check their results. This can lead to the accumulation of minor errors or inaccuracies introduced sequentially throughout data analysis. On the other hand, other analysts tend to double check their code and the results. This leads to a more branched type of data analysis results, where the same goal is cross-checked using different approaches and small nuances which are causing variability are more likely to be surfaced.</p>

Organizing model

In line with the design rationale approach, we further grouped the above categories into four major meta-categories based on their function in the model of cognitive processes involved in data analysis we propose here (Figure S9-6).

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 is where the first source of variability can be observed due to individual 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.

The interplay between the first two meta-categories is referred to by Grolemond and Wickham (2014) as the interaction between mental models and given data. Throughout the process of studying and understanding the data, an analyst updates her prior beliefs and biases with regards to what was expected vs. what is actually reflected in the data. Sometimes these discrepancies lead to updated beliefs, while in some cases an analyst internally generates an alternative explanation for the observed mismatch and rejects an alternative state of belief. This process is to some extent similar to the statistical hypothesis testing where the alternative hypothesis is either accepted or rejected. The difference is that in this case it occurs in the analyst's mind and the process is not well understood. An example for this could be a certain (perceived) understanding resulting from a professional background or personal experiences which is challenged by the data, and therefore calls these a-priori understandings into question.

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 prior to collecting or obtaining the dataset (Wagenmakers et al., 2012).

As an example of the present distinction between exploratory and confirmatory analyses, suppose that an analyst wants to find out the relation between two variables of interest. She therefore applies different methods (e.g., runs a correlation or plots different diagrams), in order to understand this relationship on a subset of the data (EDA). Once the analyst seems to have understood the meaning of these variables (i.e., has made sense of the data/problem), she wants to confirm her insights and fits a linear model on another subset of the data (CDA). At some points during the data analysis, the investigator might reach insights which interact with her personal understanding of the problem and broader system of beliefs (i.e., the cognitive sensemaking process).

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 his or her 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.

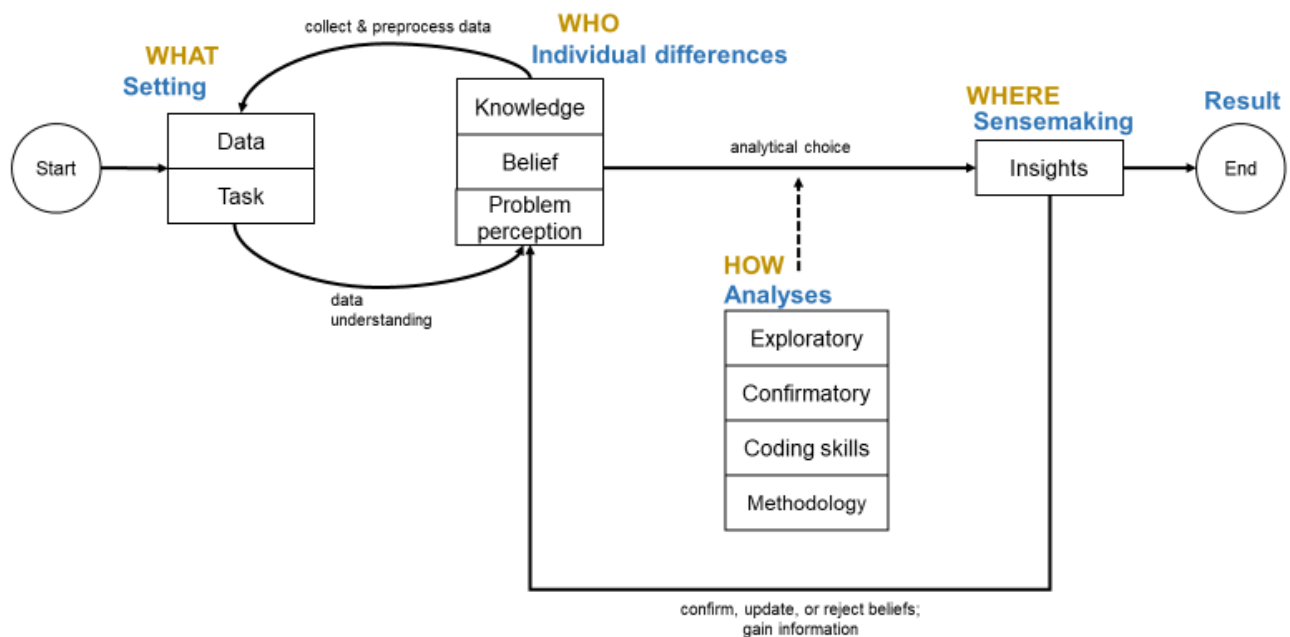


Figure S9-6. Model of a data analyst's workflow and reasoning process.

Summary

By means of the DataExplained platform and subsequent qualitative analysis, we examined which factors contribute to variability in approaches when different researchers analyzed the same data to test the same hypotheses. We also propose a model describing the process through which analysts reached their conclusions. This model draws on sensemaking theory and extends the conceptual model of data analysis proposed by Grolemond and Wickham (2014) by outlining how person factors and task settings interact to drive variability in approaches and outcomes (Figure S9-6). The proposed model 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. Crowdsourcing data analysis is not feasible for all projects, and integrating the DataExplained approach into individual or small-teams projects can help make transparent the subjective choices made during the research process.

References for Supplement 9

- Alasuutari, P. (2010). The rise and relevance of qualitative research. *International Journal of Social Research Methodology*, 13(2), 139-155.
- Barnes, W. H. (1944). The nature of explanation. *Nature*, 153(3890), 605.
- Bollier, D., & Firestone, C. M. (2010). *The promise and peril of big data* (pp. 1-6). Washington, DC: Aspen Institute, Communications and Society Program.
- 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.
- Chi, M. T. (2009). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In Vosniadou (Eds.), *International Handbook of Research on Conceptual Change* (pp. 89-110). London: Routledge.
- Conklin, E. J., & Yakemovic, K. B. (1991). A process-oriented approach to design rationale. *Human-Computer Interaction*, 6(3-4), 357-391.
- Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (pp. 146-166). Upper Saddle River, NJ: Prentice Hall.
- Fahy, P. J. (2001). Addressing some common problems in transcript analysis. *The International Review of Research in Open and Distributed Learning*, 1(2), 1-6.
- Feldman, M. (2018). *Crowdsourcing data analysis: empowering non-experts to conduct data analysis*. Unpublished dissertation, University of Zurich.
- Fox, P., & Hendler, J. (2011). Changing the equation on scientific data visualization. *Science*, 331(6018), 705-708.
- Friedkin, N. E., Proskurnikov, A. V., Tempo, R., & Parsegov, S. E. (2016). Network science on belief system dynamics under logic constraints. *Science*, 354(6310), 321-326.
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6), 460.

- Glaser, B. G., & Strauss, A.L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. London: Wiedenfeld and Nicholson.
- Grolemund, G., & Wickham, H. (2014). A cognitive interpretation of data analysis. *International Statistical Review*, 82(2), 184-204.
- Gruber, T. R., & Russell, D. M. (1992). Generative design rationale: Beyond the record and replay paradigm. In Moran and Carroll (Eds.), *Design rationale: concepts, techniques, and use*, pp. 323-349. Boca Raton: CRC Press.
- Guindon, R. (1990). Knowledge exploited by experts during software system design. *International Journal of Man-Machine Studies*, 33(3), 279-304.
- Hill, R. C., & Levenhagen, M. (1995). Metaphors and mental models: Sensemaking and sensegiving in innovative and entrepreneurial activities. *Journal of Management*, 21(6), 1057-1074.
- Hruschka, D. J., Schwartz, D., Cobb St. John, D. C., Picone-Decaro, E., Jenkins, R. A., & Carey, J.W. (2004). Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field Methods*, 16(3), 307-331.
- Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science*, 4(1), 71-115.
- Klein, G., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent systems*, 21(5), 88-92.
- Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3), 411-433.
- Kurasaki, K. S. (2000). Intercoder reliability for validating conclusions drawn from open-ended interview data. *Field Methods*, 12(3), 179-194.
- Lee, J., & Lai, K. Y. (1991). What's in design rationale? *Human-Computer Interaction*, 6(3), 251-280.
- MacLean, A., Young, R. M., Bellotti, V. M., & Moran, T. P. (1991). Questions, options, and criteria: Elements of design space analysis. *Human-Computer Interaction*, 6(3-4), 201-250.
- Miles, M. B., Huberman, A. M., & Saldana, J. (2014). Fundamentals of qualitative data analysis. In *Qualitative data analysis: A methods sourcebook*, pp. 69-104. Thousand Oaks: Sage.
- Morton, K., Balazinska, M., Grossman, D., & Mackinlay, J. (2014). Support the data enthusiast: Challenges for next-generation data-analysis systems. *Proceedings of the VLDB Endowment*, 7(6), 453-456.
- Norman, D. A. (1983). Some observations on mental models. In Gentner & Stevens (Eds), *Mental Models*. pp. 7-14. New Jersey: Lawrence Erlbaum Associates Inc.
- 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.
- Partington, D. (Ed.). (2002). *Essential skills for management research*. Thousand Oaks: Sage.
- Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993, May). The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on human factors*

- in computing systems* (pp. 269-276). Association for Computing Machinery.
- Saldana, J. (2011). *Fundamentals of Qualitative Research: Understanding Qualitative Research*. In Beretvas & Leavy (Eds.). New York: Oxford University Press.
- Schubanz, M. (2014, June). Design rationale capture in software architecture: what has to be captured? In *Proceedings of the 19th international doctoral symposium on Components and architecture* (pp. 31-36). Association for Computing Machinery.
- Seel, N. M. (2001). Epistemology, situated cognition, and mental models: 'Like a bridge over troubled water'. *Instructional Science*, 29(4-5), 403-427.
- Staub, N. (2017). *Revealing the inherent variability in data analysis*. Unpublished master's thesis, University of Zurich. DOI: 10.13140/RG.2.2.25745.53609
- Strauss, A., & Corbin, J. (1990). Basics of qualitative research: grounded theory procedure and techniques. *Qualitative Sociology*, 13(1), 3-21.
- Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2), 237-246.
- Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Boston: Addison-Wesley Publishing Company.
- 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.
- 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.
- Weiss, G., & Wodak, R. (2003). Introduction: Theory, interdisciplinarity and critical discourse analysis. In *Critical discourse analysis* (pp. 1-32). Palgrave Macmillan, London.

Appendix S9: Technical documentation for DataExplained

Below is technical documentation of the DataExplained. It should serve as an overview of the architecture as well as a guideline for setting up the necessary infrastructure.

Architecture

DataExplained is a web-based application which is built on the MEAN stack. MEAN is a JavaScript software stack used for building dynamic web applications. It builds on the components of MongoDB, Express.js, Angular and Node.js. The application is run on an Amazon EC2 instance with RStudio Server (<https://www.rstudio.com/products/rstudio/download-server/>) installed.

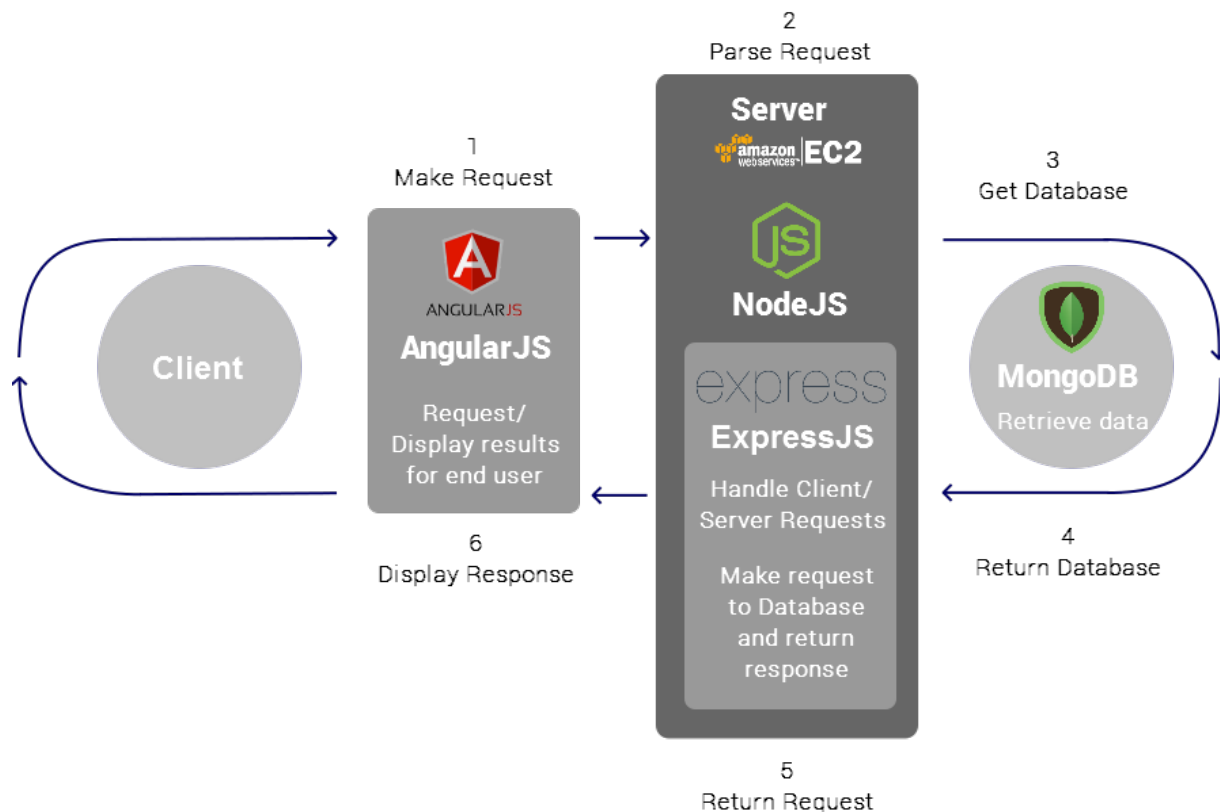


Figure S9-7. Architecture of DataExplained; figure based on Team In India (2017)

An major advantage of the MEAN stack is that both the client and the server are written in JavaScript (also known as full-stack JavaScript application). Javascript objects can easily be transformed to JSON objects, which can easily be persisted in mongoDB.

DataExplained makes use Grunt (build tool), Bower (package manager for web dependences), and NPM (package manager for nodejs dependencies).

Technical Setup

This section explains how the remote EC2 server instance is created and the setup. Additionally, instructions for the configurations on the local (developer's) machine are provided, in order to connect and commit changes to the server. Commands executed on the server are preceded with a '\$' sign. Instructions for the local machines are given for Windows systems (Windows 10). Respective configurations for other operation systems may differ. Please note that links for websites of different components may have changed.

Setup EC2 instance

In a first step, a new EC2 instance is created on aws.amazon.com. As the operation system we chose Ubuntu (Ubuntu Server 16.04 LTS). For performance reasons, the respective region where

the instance is hosted can be chosen. For DataExplained we selected “EU (Ireland)”. After the instance was created it is listed in the overview, illustrated in Figure S9-8.

<input type="checkbox"/>	Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS (IPv4)	IPv4 Public IP
<input type="checkbox"/>	dataexplained-...	i-05189db5906c50bf6	m4.large	eu-west-1a	running	2/2 checks ...	OK	ec2-34-253-169-17 eu-...	34.253.169.17

Figure S9-8. EC2 instance

During the setup, a key of the instance (e.g., “dataexplained.pem”) gets generated. Save it on your local machine under `%HOME%/.ssh/dataexplained.pem`. This key serves to connect to the instance via SSH. For this, and in order to install external packages on the server in a later step, we have to modify the instance’s security group in the AWS Management Console. For the respective rules, please see Figure S9-9 and Figure S9-10.

Security Group: sg-8effe6e8

Description Inbound Outbound Tags

Edit

Type ⓘ	Protocol ⓘ	Port Range ⓘ	Source ⓘ
HTTP	TCP	80	0.0.0.0/0
HTTP	TCP	80	::/0
Custom TCP Rule	TCP	8888	0.0.0.0/0
Custom TCP Rule	TCP	8888	::/0
Custom TCP Rule	TCP	8000	0.0.0.0/0
Custom TCP Rule	TCP	8000	::/0
SSH	TCP	22	0.0.0.0/0
Custom TCP Rule	TCP	8787	0.0.0.0/0
Custom TCP Rule	TCP	8787	::/0
Custom TCP Rule	TCP	27017	0.0.0.0/0
Custom TCP Rule	TCP	27017	::/0

Figure S9-9: EC2 instance Inbound Security Rules.

Security Group: sg-8effe6e8

Description Inbound **Outbound** Tags

Edit

Type ⓘ	Protocol ⓘ	Port Range ⓘ	Destination ⓘ
HTTP	TCP	80	0.0.0.0/0
HTTP	TCP	80	::/0
All traffic	All	All	0.0.0.0/0
All traffic	All	All	::/0

Figure S9-10: EC2 instance Outbound Security Rules

Connect via SSH

To connect via ssh, use the following command (replace path accordingly):

```
ssh -i /path/to/key/dataexplained.pem ubuntu@<your-host-name>
```

In order to make things easier in future, the following settings have to be made:

Local:

Create/update the config file under `%HOME%/.ssh/config` (with respective hostname of your EC2-instance):

Host: dataexplained

Hostname: `<your-host-name>` (e.g., `ec2-34-249-31-191.eu-west-1.compute.amazonaws.com`)

User: ubuntu

IdentityFile: `./ssh/dataexplained.pem`

EC2:

Add your personal public ssh-key of local machine on the server:

```
$ cd /.ssh
```

```
$ nano authorized_keys
```

As of now, you can connect to our remote EC2 instance via terminal:

```
ssh dataexplained
```

Configure EC2

Connect to the server via ssh and enter the following prompts:

```
$ sudo apt-get update
$ sudo apt-get install -y python-software-properties python g++ make
$ curl -sL https://deb.nodesource.com/setup_7.x | sudo -E bash -
$ sudo apt-get update
$ sudo apt-get install nodejs
$ sudo apt-get install build-essential
$ sudo apt-get install git
$ sudo apt-get install npm
$ sudo npm install cross-spawn
$ sudo npm install forever -g
$ sudo npm install pm2 -g
```

Create a bare Git repository on the server (REPO_NAME is the name for the repository you want to use):

```
$ cd /
$ mkdir REPO_NAME
$ cd REPO_NAME
$ git init -bare
```

Create a post-receive git-hook which automatically restarts the server once a new version was committed:

```
$ cd REPO_NAME/hooks/
$ touch post-receive
$ chmod +x post-receive
$ nano post-receive
```

Paste the following content:

```
#!/bin/sh
GIT_WORK_TREE=/home/ubuntu/www
export GIT_WORK_TREE
git checkout -f
```

```
cd $HOME/www
./start.sh
```

Create directory for applications content and create start script:

```
$ cd /
$ mkdir www/
$ cd /www
$ touch start.sh
$ chmod +x start.sh
$ nano start.sh
```

Paste the following content:

```
# this file is execute by post-receive hook every time a Git commit is made:
pm2 kill
export GITHUB_USER=<your github username here>
export GITHUB_SECRET=<your github password here>
export GITHUB_TOKEN=<your github token here>
sudo service mongod start
pm2 start apps.json
sudo chmod -R 777 /home/ubuntu/.pm2
```

Redirect all traffic from port 80 to 8080:

(This command has to be re-executed everytime the server is shut down or restarted!)

```
sudo iptables -t nat -A PREROUTING -p tcp -dport 80 -j REDIRECT --to-ports 8080
```

As the remote Git-repository is now configured, we need to add it on the client-side (local machine) configuration.

Create git repository in distribution folder of the application (dist) and add/edit the “config” file (within the newly created “.git” folder): `git init`

Paste the following content in the config file:

```
[remote "AWS_production"]
url = ssh://ubuntu@YOUR-IP/home/ubuntu/REPO_NAME/
fetch = +refs/heads/*:refs/remotes/REPO_NAME/*
puttykeyfile = C:\Users\YOUR-USER\.ssh\dataexplained.pem
```

From now on, the client can commit and push changes to the remote EC2 instance. This will trigger the post-receive hook, moves the application's content in the server application's folder (www), and restarts the server.

Attention: If new node-packages are added to the application (in the packages.json), you have to manually run "sudo npm install" in the /www directory.

If the application makes use of environment variables, you may consider to permanently add them to the server in order to access them (even if the start-up script would fail for some reason).

On the EC2-instance, the global environment variables are stored in `/etc/environment`. The file can be edited with `"sudo nano /etc/environment"`.

To see the newly created variables, you have to reconnect the machine via ssh and run `printenv`.

Setup MongoDB

The tutorial for installing mongoDB on an ubuntu machine (our EC2 instance) can be found on: <https://docs.mongodb.com/manual/tutorial/install-mongodb-on-ubuntu/>. It consists merely of the following steps (please refer to the link to guarantee the newest version gets installed):

```
$ sudo apt-key adv --keyserver hkp://keyserver.ubuntu.com:80 --recv
0C49F3730359A14518585931BC711F9BA15703C6

$ echo "deb [ arch=amd64,arm64 ] http://repo.mongodb.org/apt/ubuntu
xenial/mongodb-org/3.4 multiverse" | sudo tee /etc/apt/sources.list.d/mongodb-
org-3.4.list

$ sudo apt-get update

$ sudo apt-get install -y mongodb-org
```

To start mongoDB we run:

```
$ sudo service mongod start
```

To verify that mongoDB is running and to check the logs, we can inspect the contents of `"/var/log/mongodb/mongod.log"`. The running port is configured in `"/etc/mongod.conf"` and is set to 27017 by default.

Setup RStudio Server

The tutorial on how to install RStudio server can be found on: <https://www.rstudio.com/products/rstudio/download-server/>. It consists merely of the following steps (please refer to the link to guarantee the newest version gets installed):

```
$ sudo apt-get install r-base
$ sudo apt-get install gdebi-core
$ wget https://download2.rstudio.org/rstudio-server-1.0.136-amd64.deb
$ sudo gdebi rstudio-server-1.0.136-amd64.deb
```

To allow RStudio Server to run in an iframe, you have to add the following configuration in `"/etc/rstudio/rserver.conf"`:

```
www-frame-origin=anyline
```

For the sessions in RStudio its beneficial to enable automatic saving of the workspace and set the default workspace environment. You can do so by adding following the following configuration in `/etc/rstudio/rsession.conf"`:

```
session-save-action-default=yes session-default-working-dir= /rstudio-workspace
```

To manually stop, start, and restart the server you use the following commands:

```
$ sudo rstudio-server stop
$ sudo rstudio-server stop
$ sudo rstudio-server restart
```

Each time you change a configuration file, you have to either restart Rstudio Server with the commands listed above, or you can run:

```
$ sudo rstudio-server verify-installation
```

To add R-packages globally (available for all users in their workspace without prior installation) you can follow these steps (example shown for package "readr"):

```
$ cd
$ sudo wget https://cran.r-project.org/src/contrib/readr_1.0.0.tar.gz
$ sudo R CMD INSTALL -l /usr/lib/R/library readr_1.0.0.tar.gz
$ sudo rm -r readr_1.0.0.tar.gz
```

If the package has dependencies of other packages which are not installed on the server yet, you need to install these packages first.

Run, Build, Deploy

This section serves as a guidance to locally run DataExplained for development, as well as build and deploy a new version to the server.

Prerequisites

- Node.js and npm (Node 4.2.3, npm 2.14.7)
- Bower (`npm install --global bower`)
- Grunt (`npm install --global grunt-cli`)
- MongoDB daemon running (default port 27017) with mongod

Development

1. Run `npm install` to install server dependencies.
2. Run `bower install` to install front-end dependencies.
3. Run `mongod` in a separate shell to keep an instance of the MongoDB Daemon running.
4. Run `grunt serve` to start the development server. It should automatically open the client in your browser when ready.

Deployment

1. Run `grunt build` for building the application. The built application is then contained in the "dist" folder.
2. Make sure that the newly created files under *dist/client/app* are all added to git.

3. Push the changes to the remote git repository on the server (located under `~/REPO_NAME` on your server instance).
4. Due to the configured git-hook on the remote repository, the server automatically replaces the application content and restarts. This does not take more than a few seconds.
5. Additionally you may want to push the changes to the git repository of DataExplained.

Database Backup

To backup the database you have to run the following command:

```
$ mongodump -out /home/ubuntu/backup/
```

You can restore your backed up dump with ("dataexplained" refers to the name of the database):

```
$ mongorestore -d dataexplained /home/ubuntu/backup/
```

Cronjobs

In order to periodically run jobs on the server (i.e., backup the database), you can define cronjobs by executing

```
$ crontab -e
```

The following cronjobs are recommended for DataExplained:

- Hourly backup of the database via `mongo_backup.sh` script. (Note, this script additionally uploads the compressed backup to Amazon S3.)

```
$ 00 0-23 * * * /bin/bash /home/ubuntu/backup/mongo_backup.sh
```

- Remove database backups on server older than 7 days to save resources. Executed once a day.

```
$ 01 05 * * * /usr/bin/find /home/ubuntu/backup/rationalecap/ -mtime +7 -  
exec rm \;
```

- Run script which checks whether the database is connected or not. If not, the script will reconnect it. Executed every five minutes.

```
$ */5 * * * * /bin/bash /home/ubuntu/www/mongocheck.sh >/dev/null 2>&1
```

- Send metrics to Amazon AWS which can be fetched via the CloudWatch service. Executed every half-hourly.

```
$ 30 * * * * /aws-scripts-mon/mon-put-instance-data.pl -disk-path=-  
disk-space-util -disk-space-used -disk-space-avail -from-cron
```

Reference for Appendix S9

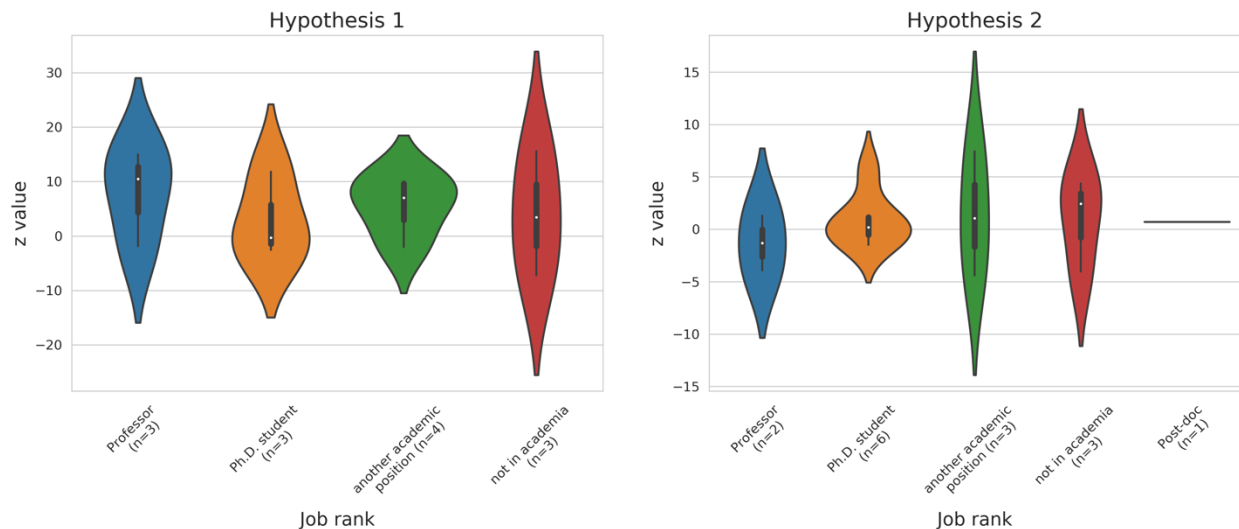
Team In India (2017). MEAN Stack Components. Retrieved from
<https://www.teaminindia.com/hire-mean-stack-developer.html>

Supplement 10: Exploratory analysis on analyst expertise and effect size dispersion

We conducted an internal exploratory analysis to investigate the extent to which effect size dispersion might be driven by either junior or more senior analysts. We conclude from this analysis that job rank does not seem to affect either the magnitude or the direction of estimates.

The violin plots below report the results from our exploratory analysis. The y-axis indicates effect sizes found by analysts as z-values and the x-axis indicates analysts' job rank, from professors (in blue) on the left hand side to analysts not in academia (in red) on the right hand side. Numbers in brackets after job titles indicate the number of analysts for each job category.

White dots indicate median values, the thick part of the gray bar indicates the interquartile range, and the thin part of the gray bar indicates the rest of the distribution. The shape of each violin plot is given by a kernel density estimation that indicates how the data is distributed. Wide parts of a plot indicate a higher probability for this effect size to be found and thin parts of each plot indicate a lower probability for this effect size to be found.



The figures only include analyses deemed error-free. One very large z-value (=106.27) for H1 Analyst 4 (not in academia) has been excluded to enhance the readability of the violin plots. The dataset and code (as a Jupyter notebook) used to create these violin plots are available on the Open Science Framework here: <https://osf.io/k5uj6/>.

Supplement 11: Further details on the Boba multiverse analyses

We conducted the Boba multiverse analyses on 2,977 universes for H1 and 14,835 universes for H2 (see Table 11-1 for details on this).

Besides visualizing the overall z-score distribution of all universes, we also examine the trends and patterns of z-scores from different alternative analytic approaches within each branch.

To do so, we use a trellis plot, where each subplot depicts the subset of universes that adopt a particular analytic alternative. In all trellis plots, the x-axis represents the magnitude of the z-score, and the y-axis represents count. Negative z-scores are highlighted in red. The trellis plot allows a richer understanding of the sensitivity of a branch. If a branch is not sensitive, we would expect to see roughly the same distributions across all subplots.

Figures S11-1 through S11-8 show the top four most sensitive branches in H1 and H2, respectively, according to the k-samples Anderson Darling test (see Boba multiverse section of the Results in the main manuscript). In Figure S11-1, we can see that different DV operationalizations correspond to very different z-score distributions. For example, every universe with *Female_Contributions* as the DV would only produce positive z-scores, while choosing *WC_ContributionsbyAuthor* or *NumPosts* as the DV would result in plenty of negative estimates. In contrast, Figure S11-4 shows that different choices of models do not correspond to wildly different z-score distributions, suggesting that model is not a particularly sensitive branch. These observations agree with the standardized test statistics in Table 5 of the main manuscript.

Similarly, we might observe how different analytic approaches influence the estimates in H2. While the overall z-score distribution is quite symmetrical around zero (Figure 7 in the main manuscript), some IV operationalizations produce a majority of negative z-scores, such as *AcademicHierarchy*, while others tend to produce positive estimates, such as *PhdRanking* (Figure S11-5).

Hypothesis	Full cross-product	Excluded invalid combinations	Excluded run-time errors
H1	5,511,240	2,984	2,977
H2	13,608,000	15,257	14,835

Table 11-1. *Size of the H1 and H2 multiverses. From left to right, the columns show the full cross-product of all analytic choices, the remaining universes after excluding invalid combinations, and the remaining universes after further excluding run-time errors.*

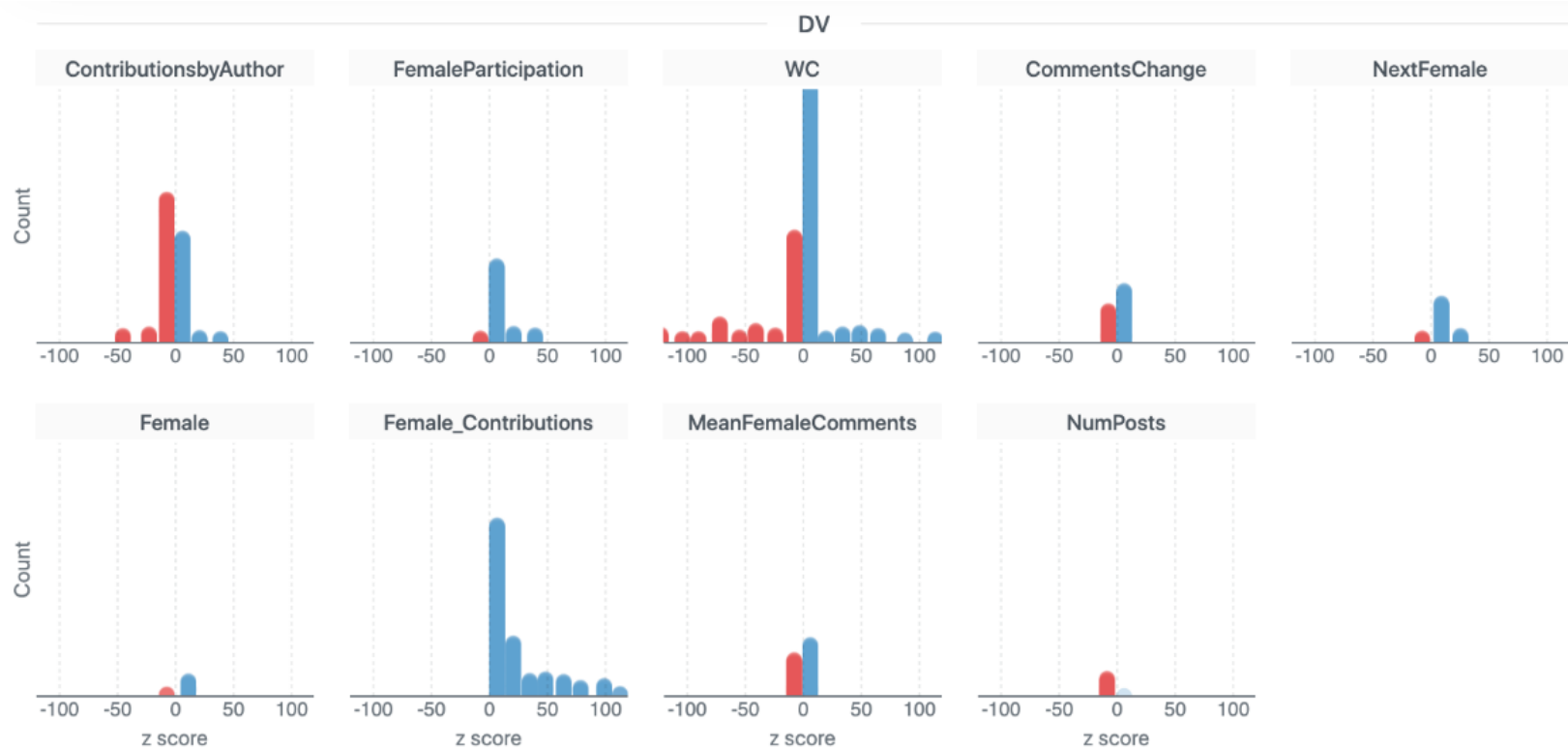


Figure S11-1. Trellis plot of the most sensitive branch in H1 – different operationalizations of the dependent variable (DV).

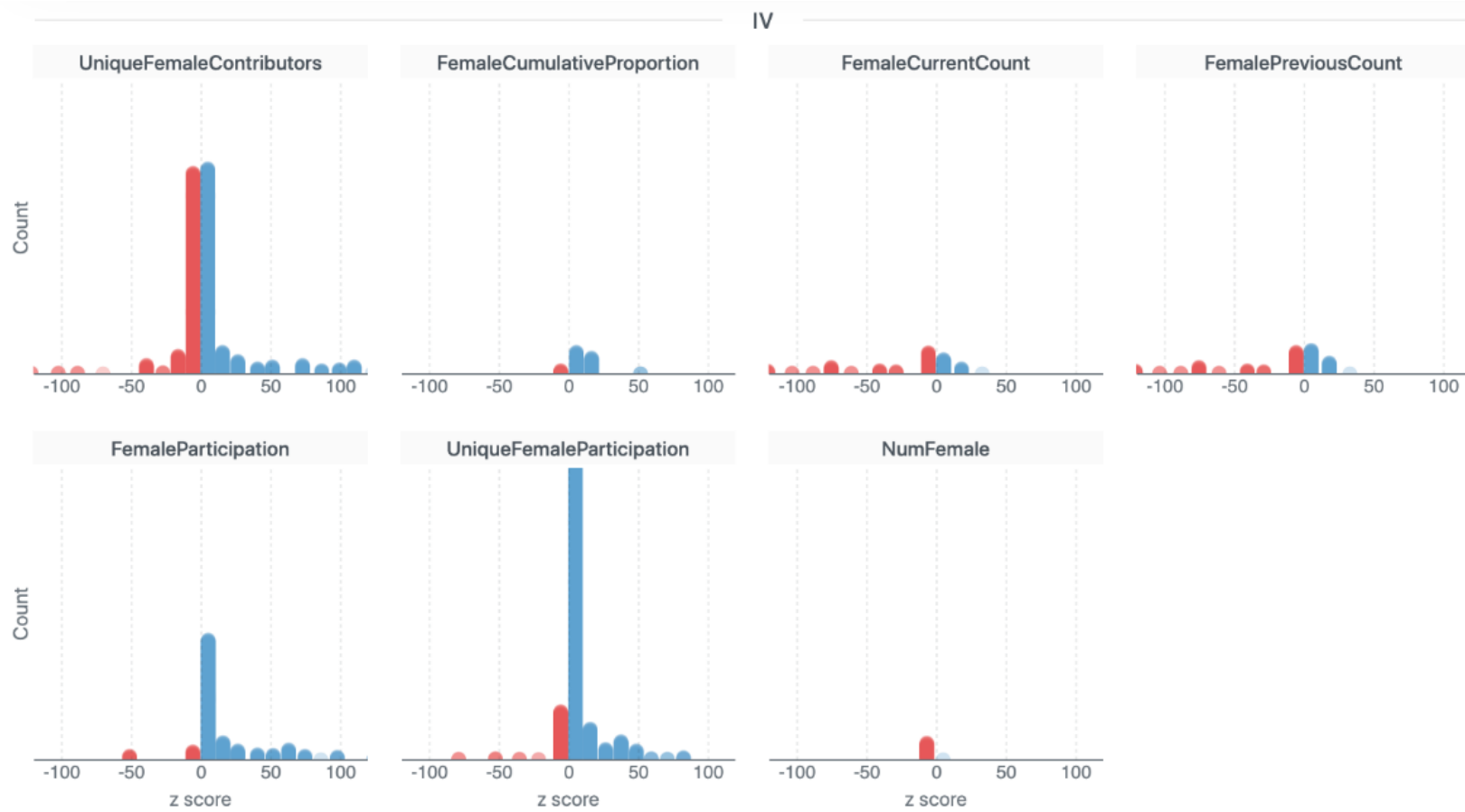


Figure S11-2. Trellis plot of the second most sensitive branches in H1 – different operationalizations of the independent variable (IV)



Figure S11-3. Trellis plot of the third most sensitive branches in H1 – different choices of the unit of analysis.

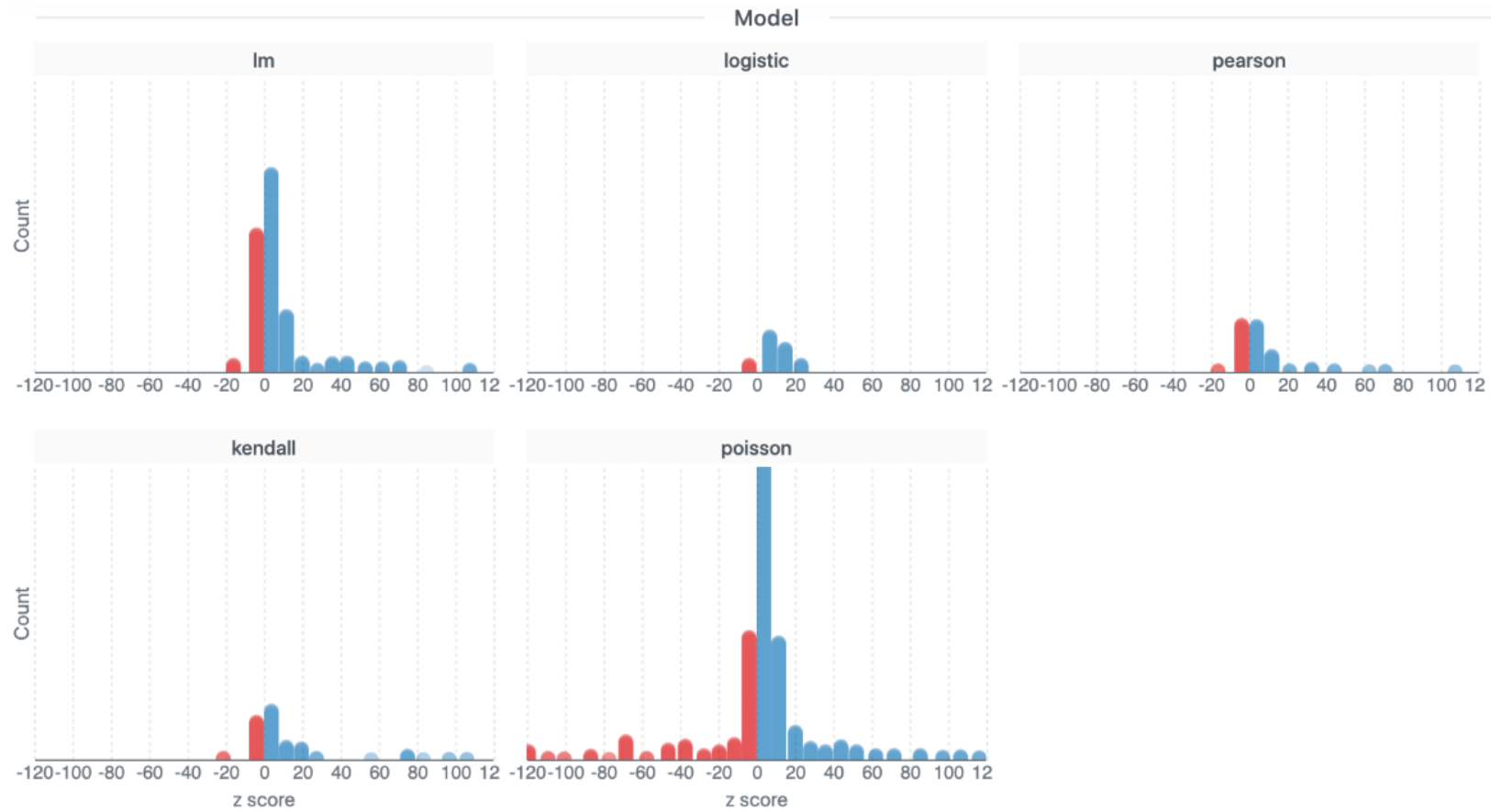


Figure S11-4. Trellis plot of the fourth most sensitive branches in H1 – different choices of the statistical model.

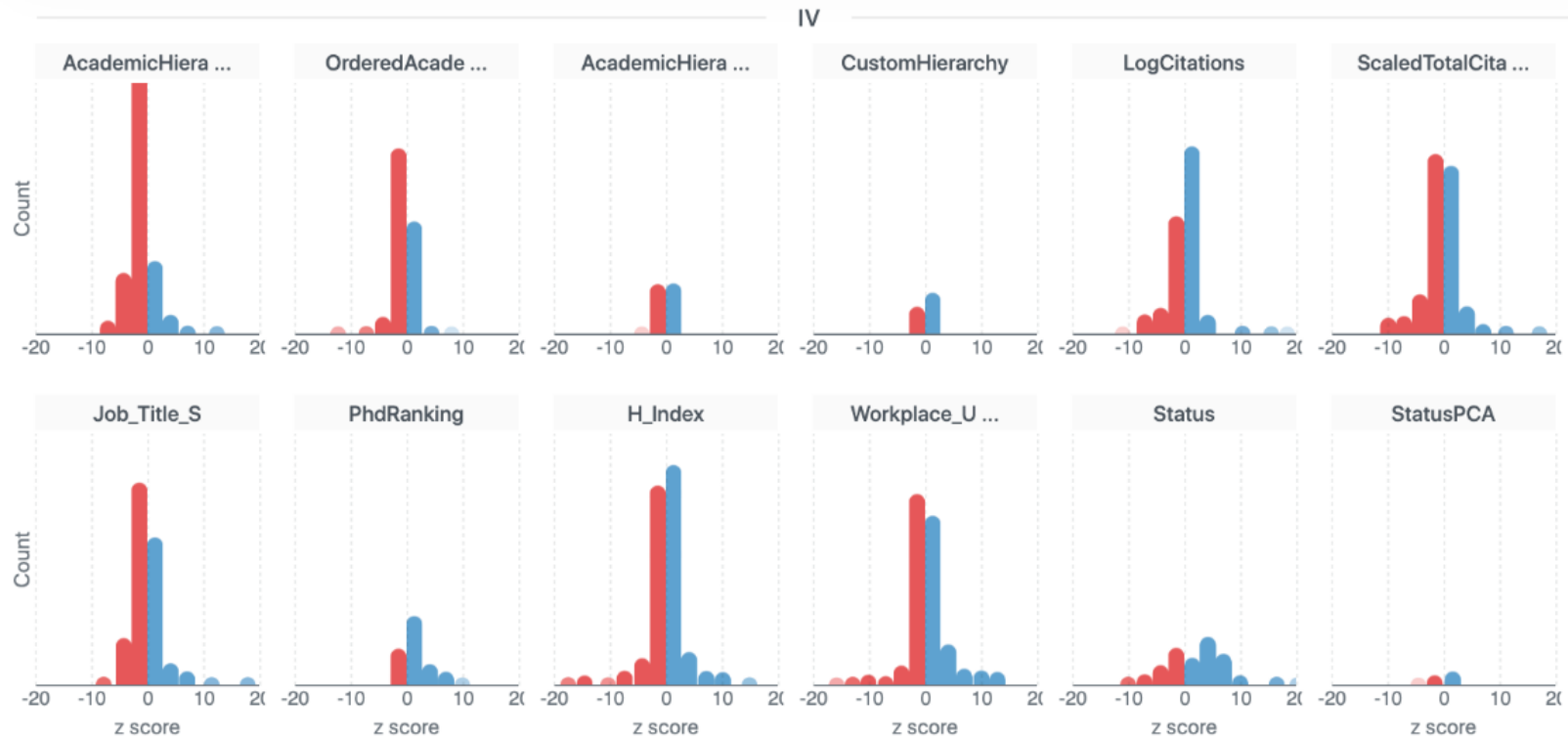


Figure S11-5. Trellis plot of the most sensitive branch in H2 – different operationalizations of the independent variable (IV).

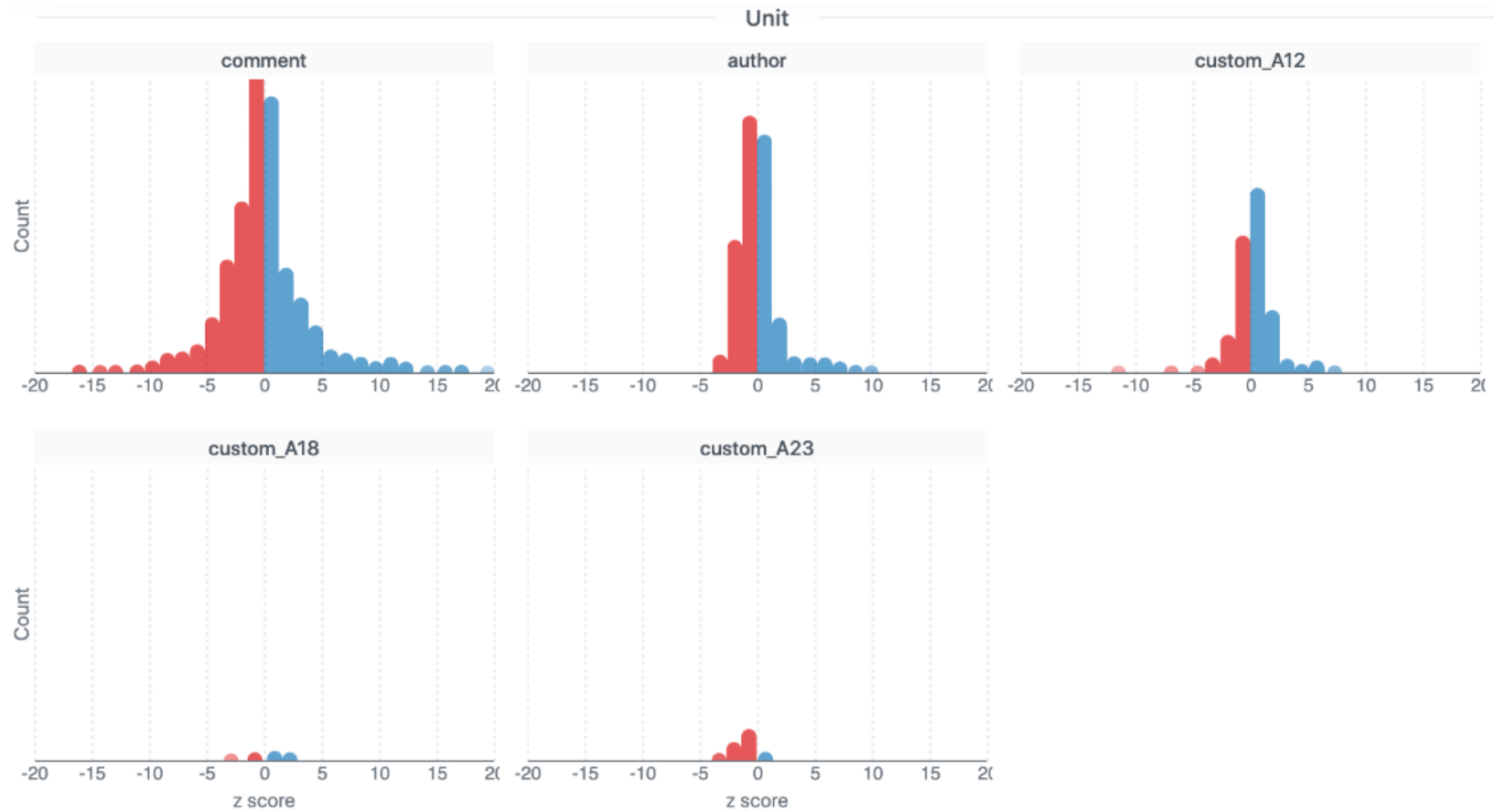


Figure S11-6. Trellis plot of the second most sensitive branch in H2 – different choices of the unit of analysis.

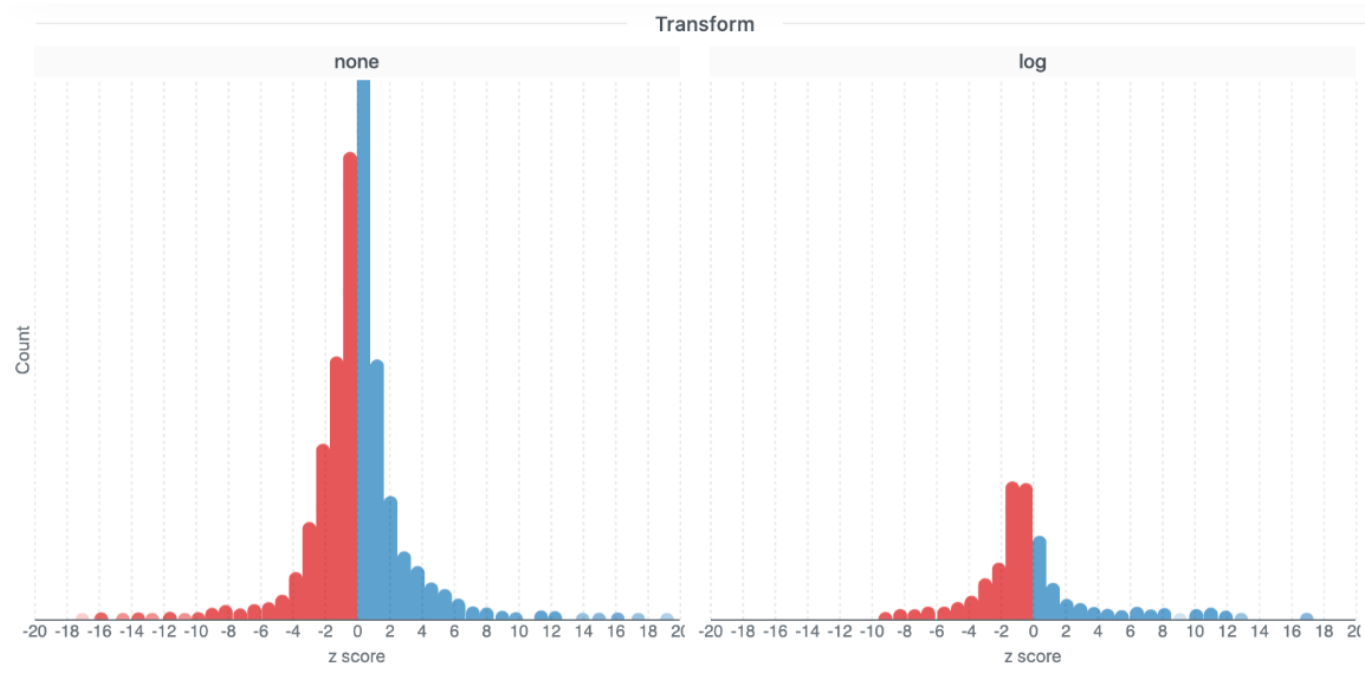


Figure S11-7. Trellis plot of the third most sensitive branch in H2 – whether the dependent variable is log-transformed.

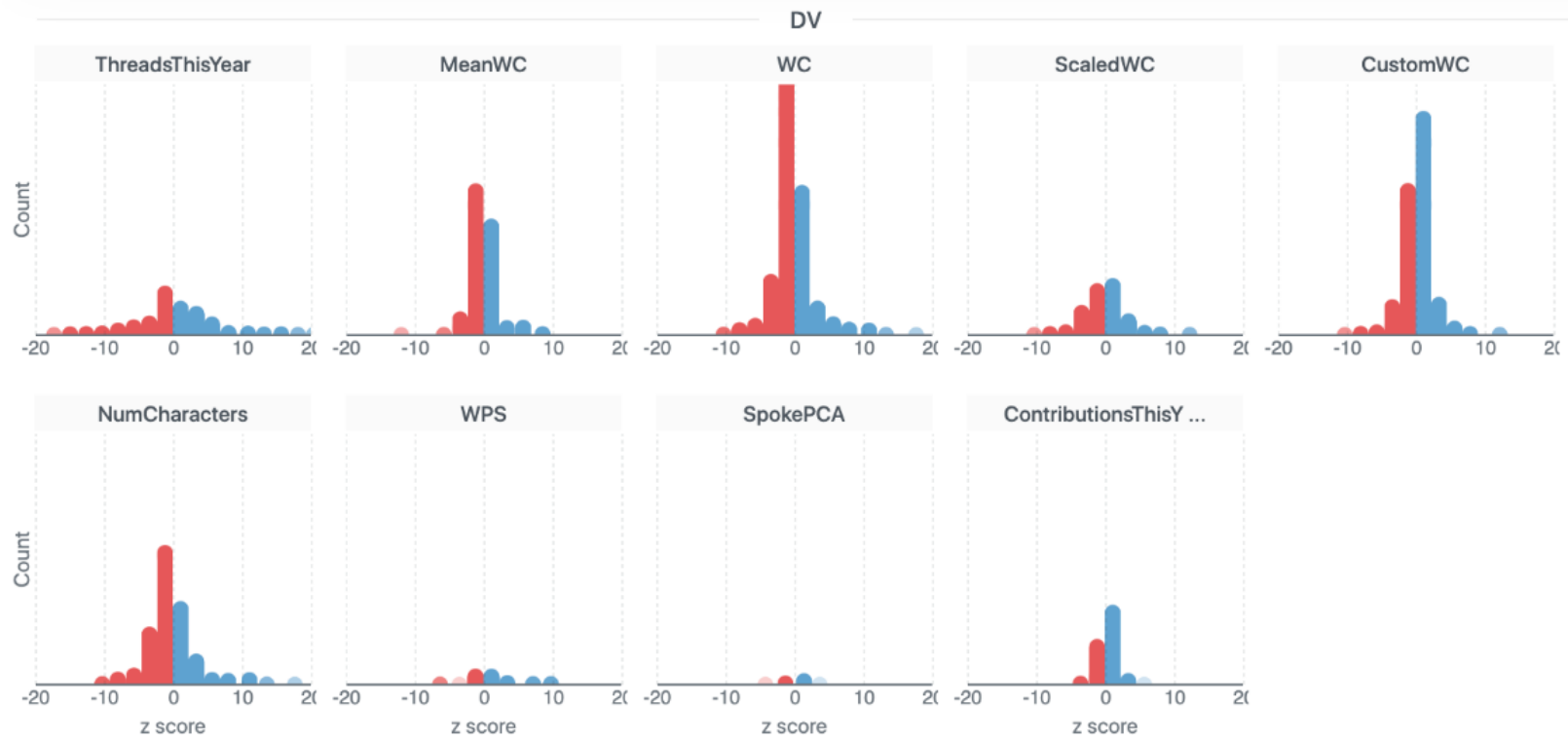


Figure S11-8. Trellis plot of the fourth most sensitive branch in H2 – different operationalizations of the dependent variable (DV).