Uncertain Science is Bad Science: Lay Perceptions of Crowd-Scientific Findings

Shilaan Alzahawi1 & Benoît Monin1

1 Stanford University, Graduate School of Business

Author note

Correspondence concerning this article should be addressed to Shilaan Alzahawi, 655 Knight Way, Stanford, CA 94305. E-mail: [shilaan@stanford.edu](mailto:shilaan@stanford.edu)

Abstract

A recent movement towards crowd science has emerged in the behavioral sciences. Crowd science aims to increase the rigor, reliability, and credibility of scientific research. Does it meet its promises in reality? We report the results of an experiment in which we explore whether scientific findings emerging from a crowd of researchers (vs. a typical research collaboration) are more likely to sway the prior beliefs of research consumers, increase ratings of credibility and confidence, and decrease ratings of bias and error. We focus on crowdsourced data analysis (also known as the ‘many analysts’ or ‘multi-analyst’ approach): giving the same dataset to different teams of scientists, who independently analyze it to answer the same research question.  
In line with our hypotheses, we find that lay consumers of multi-analyst studies with inconsistent results (compared to single-analyst studies) are less likely to be swayed, have less confidence in the average effect size estimate, find the results less credible, and believe the results are more likely to stem from bias and error. Contrary to our hypotheses, we do not find that multi-analyst studies with consistent results (compared to single-analyst studies) increase the sway and credibility of scientific research.

*Keywords:* Meta-science, Crowd science, Many analysts, Multi-analyst, Research credibility

*Word count:* X

Uncertain Science is Bad Science: Lay Perceptions of Crowd-Scientific Findings

The credibility of scientific research is in doubt, among lay consumer and scientist alike. From man-made climate change to the safety and effectiveness of Covid-19 vaccinations, important scientific findings are rejected at large (Hornsey & Fielding, 2017). In addition to widespread science skepticism among the general public, doubts about the credibility of scientific research have emerged from within the scientific community itself. Reliance on widely followed research practices, some scientists have come to realize, can generate impossible results, such as the ability of humans to feel what’s in the future (Bem, 2011) or to become younger in age from listening to certain music (Simmons, Nelson, & Simonsohn, 2011: p. 1360).  
What can be done about this “crisis of confidence” (Pashler & Wagenmakers, 2012: p. 528)? Several tools have been proposed to improve the rigor and reliability of scientific research and, consequently, combat the crisis of confidence. One such tool is the crowd or big team science approach, which leverages a large number of individuals or teams at specific stages of the research process (Uhlmann et al., 2019).  
Crowd science aims to improve the rigor, reliability, and credibility of scientific research. Does it meet its promises in reality? In this commentary, we focus on crowdsourced data analysis (also known as the ‘many analysts’ or ‘multi-analyst’ approach): giving the same dataset to different teams of scientists, who independently analyze it to answer the same research question. We explore whether multi-analyst studies increase the sway and credibility of scientific research. That is, are scientific findings emerging from a crowd of researchers (vs. a typical science collaboration) more likely to sway the prior beliefs of research consumers, increase ratings of credibility and confidence, and decrease ratings of bias and error?

# Methods

The preregistration of our experiment can be found in the Open Science Foundation (OSF) Registries at <https://osf.io/rpu98>. All data and code needed to reproduce this article can be found on GitHub at <https://github.com/shilaan/many-analysts> and the OSF at <https://osf.io/vedb4/>. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

## Participants

We sampled participants from the Prolific participant pool. We included participants from the United States or the United Kingdom who were at least 18 years old and spoke English as a first language. Participants were paid $2 for 8-12 minutes of their time.  
Our target sample size was 1500 participants after exclusions due to failing the attention check, not completing the attention check, or attempting to take the survey more than once. We aimed to sample until reaching the desired number of valid participants or until 2 weeks passed (whichever came first). We expected relatively small effect sizes and decided on an upper limit on the number of participants based on monetary constraints.

## Procedure and materials

The Survey Materials can be found on the OSF at <https://osf.io/md9z5/>. After a brief introduction to a research question (“Do religious people report higher well-being?”), participants were first asked to report their prior beliefs (“How likely do you think it is that people who are more religious generally report higher well-being?”) on a slider from 0% (not likely at all) to 100% (extremely likely).  
After reporting their prior beliefs, participants were randomly allocated to one of three experimental conditions in which they learned about the approach and findings of a single-analyst study (the “single-analyst” condition); a multi-analyst study with consistent results (the “multi-analyst: consistent” condition); or a multi-analyst study with inconsistent results (the “multi-analyst: inconsistent” condition).  
Afterwards, participants rated (1) their final beliefs about the research question, (2) the credibility of the results, (3) their confidence in the effect size estimate, and how likely it is that the estimate was influenced by (4) bias, (5) error, and (6) degrees of freedom. All questions were answered on a slider from 0% (not likely/credible/confident at all) to 100% (extremely likely/credible/confident).

## Data analysis

For all six dependent variables, we ran linear regression models with condition as the independent variable (with the single-analyst condition as the reference category) and prior beliefs as a covariate. All statistical models and code were preregistered at <https://osf.io/rpu98>. An overview of our preregistered, directional hypotheses can be found in Table 1. We did not preregister any hypotheses for the last measure (“degrees of freedom”). Thus, the findingds concerning the impact of experimental condition on ratings of degrees of freedom are exploratory, and should be treated as such.

Table 1: Table 1

|  |  |  |
| --- | --- | --- |
| Measure | Many-analysts: Consistent | Many-analysts: Inconsistent |
| 1. Final beliefs | + | - |
| 2. Credibility | + | - |
| 3. Confidence | - | - |
| 4. Bias | - | + |
| 5. Error | - | + |

# Results

## Participants

# Discussion

## Acknowledgement

This manuscript was created using R (Version 4.0.5; R Core Team, 2021) and the R-packages *dplyr* (Version 1.0.7; Wickham, François, Henry, & Müller, 2021), *forcats* (Version 0.5.1; Wickham, 2021a), *ggplot2* (Version 3.3.5; Wickham, 2016), *here* (Version 1.0.1; Müller, 2020), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *purrr* (Version 0.3.4; Henry & Wickham, 2020), *readr* (Version 1.4.0; Wickham & Hester, 2020), *rmarkdown* (Version 2.10; Xie, Allaire, & Grolemund, 2018; Xie, Dervieux, & Riederer, 2020), *stringr* (Version 1.4.0; Wickham, 2019), *tibble* (Version 3.1.4; Müller & Wickham, 2021), *tidyr* (Version 1.1.3; Wickham, 2021b), *tidyverse* (Version 1.3.1; Wickham et al., 2019), and *tinylabels* (Version 0.2.1; Barth, 2021).

## Data availability statement

The data that support the findings of this study are openly available on GitHub at <https://github.com/shilaan/many-analysts> and the OSF at <https://osf.io/vedb4/>.

# References

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from <https://github.com/crsh/papaja>

Barth, M. (2021). *tinylabels: Lightweight variable labels*. Retrieved from <https://github.com/mariusbarth/tinylabels>

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. <https://doi.org/10.1037/a0021524>

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools*. Retrieved from <https://CRAN.R-project.org/package=purrr>

Hornsey, M. J., & Fielding, K. S. (2017). Attitude roots and Jiu Jitsu persuasion: Understanding and overcoming the motivated rejection of science. *The American Psychologist*, *72*(5), 459–473. <https://doi.org/10.1037/a0040437>

Müller, K. (2020). *Here: A simpler way to find your files*. Retrieved from <https://CRAN.R-project.org/package=here>

Müller, K., & Wickham, H. (2021). *Tibble: Simple data frames*. Retrieved from <https://CRAN.R-project.org/package=tibble>

Pashler, H., & Wagenmakers, E. (2012). Editors’ Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. <https://doi.org/10.1177/1745691612465253>

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>

Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., … Nosek, B. A. (2019). Scientific Utopia III: Crowdsourcing Science. *Perspectives on Psychological Science*, *14*(5), 711–733. <https://doi.org/10.1177/1745691619850561>

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <https://ggplot2.tidyverse.org>

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from <https://CRAN.R-project.org/package=stringr>

Wickham, H. (2021a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from <https://CRAN.R-project.org/package=forcats>

Wickham, H. (2021b). *Tidyr: Tidy messy data*. Retrieved from <https://CRAN.R-project.org/package=tidyr>

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. <https://doi.org/10.21105/joss.01686>

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation*. Retrieved from <https://CRAN.R-project.org/package=dplyr>

Wickham, H., & Hester, J. (2020). *Readr: Read rectangular text data*. Retrieved from <https://CRAN.R-project.org/package=readr>

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from <https://bookdown.org/yihui/rmarkdown>

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R markdown cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from <https://bookdown.org/yihui/rmarkdown-cookbook>