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, find the results less credible, have less confidence in the average effect size estimate, 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 to lay consumer.

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

*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. What can be done about this “crisis of confidence” (Pashler & Wagenmakers, 2012: p. 528)? Several tools have been proposed to improve the reliability of scientific research and combat the crisis of confidence. One such tool is the crowd 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? Here, 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 to lay consumers. 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. We excluded participants who failed the attention check or attempted to take the survey more than once. Our target sample size was 1500 participants after exclusions. We aimed to keep sampling 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

All 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 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 scientific study. In the single-analyst condition, a single team of six researchers reports a 5% increase in well-being among religious people; in the multi-consistent condition, six independent researchers report six consistent estimates (2%, 4%, 5%, 5%, and 6%, respectively) that average to 5%; and in the multi-inconsistent condition, six independent researchers report six inconsistent estimates (-6%, -2%, 5%, 5%, 12%, and 16%, respectively) that average to 5%.  
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 measures, we run 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 hypotheses, 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; the findings concerning the impact of experimental condition on ratings of degrees of freedom are exploratory, and should be treated as such. Because we test five separate hypotheses using two comparisons each (one comparison of the single-analyst vs. the multi-consistent condition, and one comparison of the single-analyst vs. the multi-inconsistent condition), we use the Bonferroni method to correct for multiple comparisons. Thus, our preregistered threshold for statistical significance is .

# Results

We recorded 2016 responses in Qualtrics. We excluded 116 observations from participants who attempted to take the survey more than once, 73 participants who were screened out prior to starting the survey or did not consent, and 327 participants who failed the attention check. This left us with a total sample of 1500 participants (500 per condition).

Our main findings are displayed in Figure 1. Controlling for prior beliefs and comparing to the single-analyst condition, we found that (1) reported final beliefs were significantly lower in both the multi-consistent condition, , 95% CI , , , and the multi-inconsistent condition, , 95% CI , , ; (2) ratings of credibility were significantly lower in the multi-inconsistent condition, , 95% CI , , , while they were not significantly different in the multi-consistent condition, , 95% CI , , ; (3) confidence in the effect size estimate was significantly lower in the multi-inconsistent condition, , 95% CI , , , while it was not significantly different in the multi-consistent condition, , 95% CI , , ; (4) ratings of bias were significantly greater in the multi-inconsistent condition, , 95% CI , , , while they were not significantly different in the multi-consistent condition, , 95% CI , , ; and (5) ratings of error were significantly greater in both the multi-consistent condition, , 95% CI , , , and the multi-inconsistent condition, , 95% CI , , . For our exploratory measure of discretion, we found that ratings of degrees of freedom were significantly greater in both the multi-consistent condition, , 95% CI , , , and the multi-inconsistent condition, , 95% CI , , .  
In sum, all preregistered hypotheses (see Table 1) related to the multi-consistent condition were not corroborated; all preregistered hypotheses related to the multi-inconsistent condition were corroborated. In line with our hypotheses, multi-analyst studies with inconsistent results decrease the sway and credibility of scientific research to lay consumers. Contrary to our hypotheses, multi-analyst studies with consistent result do not increase the sway and credibility of scientific research to lay consumers. Figure 2 further clarifies the sway of multi-analyst vs. single-analyst studies, by displaying the distribution of prior and final beliefs across the three conditions.

# Discussion

From the proliferation of big team science and large-scale replication initiatives to preregistration and registered reports, several scientific fields have undergone significant reform with the well-intended goal of improving the rigor and reliability of scientific research. However, as with any real-world intervention, scientific reform can have unintended consequences. Here, we focused on the effects of crowdsourcing data analysis, and find that the multi-analyst approach may have an unintended consequence. While instituted with the goal of improving the credibility of scientific research, lay consumers appear to resist the inherent uncertainty that comes with inconsistent estimates. As a result, they are less likely to believe in the reported phenomenon, rate the scientific finding as less credible, have less confidence in the reported effect size estimate, and are more likely to think that the findings stem from bias and error.

## Acknowledgements

This manuscript was created using R (Version 4.0.5; R Core Team, 2021) and the R-packages *broom* (Version 0.7.9.9000; Robinson, Hayes, & Couch, 2021), *dplyr* (Version 1.0.7; Wickham, François, Henry, & Müller, 2021), *ggplot2* (Version 3.3.5; Wickham, 2016), *glue* (Version 1.4.2; Hester, 2020), *here* (Version 1.0.1; Müller, 2020), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *purrr* (Version 0.3.4; Henry & Wickham, 2020), *raincloudplots* (Version 0.2.0; person), 2021), *readr* (Version 1.4.0; Wickham & Hester, 2020), *rmarkdown* (Version 2.10; Xie, Allaire, & Grolemund, 2018; Xie, Dervieux, & Riederer, 2020), and *tinylabels* (Version 0.2.1; Barth, 2021). We thank Nicole Clare Kolmstetter for valuable assistance in the data collection.

## 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: Prepare reproducible APA journal articles with R Markdown*. Retrieved from <https://github.com/crsh/papaja>

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

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

Hester, J. (2020). *Glue: Interpreted string literals*. Retrieved from <https://CRAN.R-project.org/package=glue>

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

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>

person). (2021). *Raincloudplots: The easy way to create raincloud plots*. Retrieved from <https://github.com/jorvlan/raincloudplots>

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/>

Robinson, D., Hayes, A., & Couch, S. (2021). *Broom: Convert statistical objects into tidy tibbles*.

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., 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>

# Tables and Figures

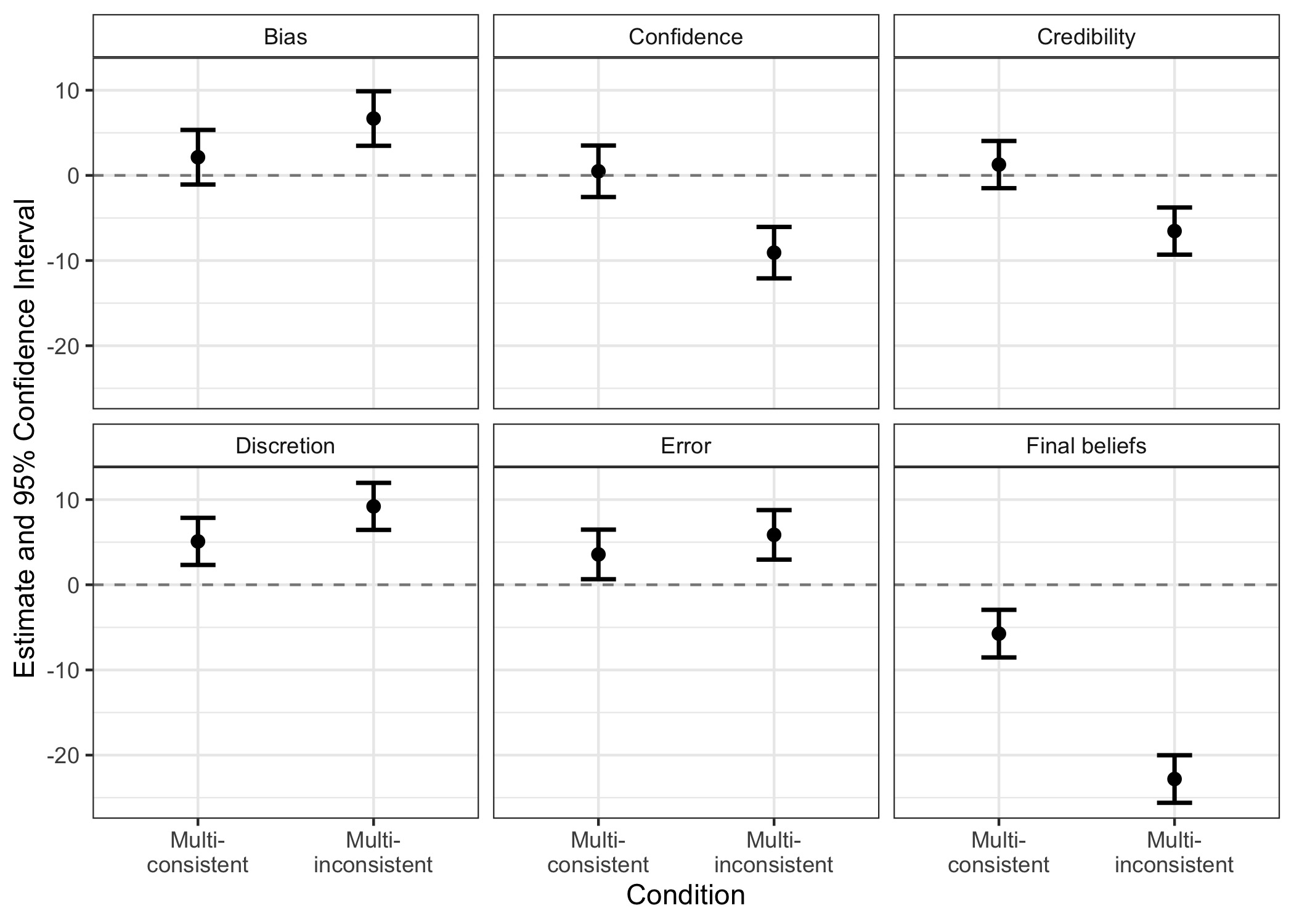
## Table 1

Predicted direction of effects for all dependent variables, compared to the single-analyst condition and controlling for prior beliefs

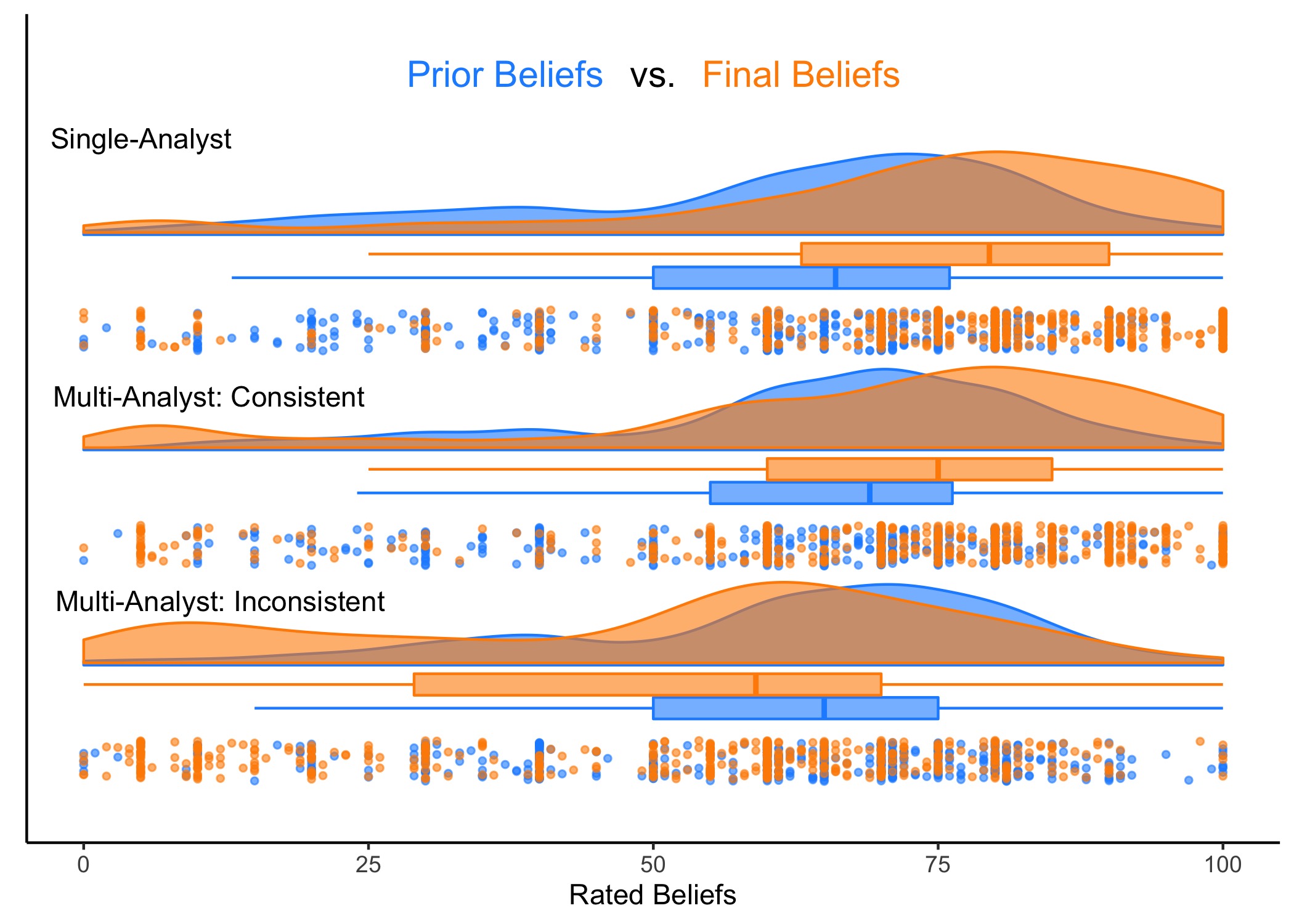
|  |  |  |
| --- | --- | --- |
| Measure | Many-analyst: Consistent | Many-analyst: Inconsistent |
| 1. Final beliefs | + | - |
| 2. Credibility | + | - |
| 3. Confidence | - | - |
| 4. Bias | - | + |
| 5. Error | - | + |

*Note*. Table 1 indicates the predicted direction of the effect for each of the five dependent variables, compared to the single-analyst condition and controlling for prior beliefs. For example, we hypothesized that, compared to a single-analyst study and controlling for prior beliefs, ratings of credibility would be greater in the multi-analyst: consistent condition and lower in the multi-analyst: inconsistent condition.

## Figure 1

*Ratings of Bias, Confidence, Credibility, Discretion, Error, and Final Beliefs*  *Note.* Figure 1 displays coefficient estimates (and 95% confidence intervals) of bias, confidence, credibility, discretion, error, and final beliefs in the two multi-analyst conditions, compared to the single-analyst condition (and controlling for prior beliefs).

## Figure 2

*Individual data points, quartiles, and distributions of prior and final beliefs in the single-analyst, multi-consistent, and multi-inconsistent conditions*  
  *Note.*  Prior beliefs are displayed in blue; final beliefs are displayed in orange. The respective boxes display the lower quartiles, medians, and upper quartiles of prior and final beliefs by condition.