Lay Perceptions of Crowd-Scientific Findings: The Risks of Variability and Lack of Consensus

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. One of the aims of crowd science is to increase the credibility of scientific research. Does it meet this promise in reality? We run an experiment to study the effects of scientific findings emerging from a crowd of researchers (vs. a typical research collaboration) on lay consumers’ posterior beliefs and ratings of credibility, confidence in an aggregate effect size estimate, 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 the data to answer a research question and/or estimate a parameter of interest. We compare the effects of providing (a) a single, aggregate parameter estimate (the single-analyst condition) vs. multiple parameter estimates that (b) vary slightly and are all positive, leading to the same qualitative conclusion (the “multi-consistent” condition) or (c) vary widely and are of both signs, leading to differing qualitative conclusions (the “multi-inconsistent” condition). In line with our hypotheses, we find that lay consumers of multi-analyst studies with inconsistent results (compared to single-analyst studies) have lower posterior beliefs, 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 findings to lay consumers: instead, to our surprise, they lead to lower posterior beliefs and higher ratings of error.

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

*Word count:* 3,627

Lay Perceptions of Crowd-Scientific Findings: The Risks of Variability and Lack of Consensus

The credibility of scientific research is in doubt, among lay consumer (Hornsey & Fielding, 2017) and scientist (Pashler & Wagenmakers, 2012) alike. Several tools have been proposed to combat this “crisis of confidence” (Ibid., p. 528). 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). Does crowd science meet its promise – to improve the credibility of scientific research – in reality?  
We explore the effects of scientific findings emerging from a crowd of researchers (vs. a typical research collaboration) on lay consumers’ posterior beliefs, perceptions of credibility, confidence in an aggregate effect size estimate, and 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 and/or estimate a parameter of interest. We compare the effects of providing (a) a single, aggregate parameter estimate (the single-analyst condition) vs. multiple parameter estimates that (b) vary slightly and are all positive, leading to the same qualitative conclusion (the “multi-consistent” condition) or (c) vary widely and are of both signs, leading to differing qualitative conclusions (the “multi-inconsistent” condition).  
According to previous work, crowdsourced data analysis should increase the credibility of scientific findings; Table 1 provides an overview of the proposed confidence- and credibility-related benefits of the multi-analyst approach. Normatively, we expect laypeople who observe several scientists independently come to the same qualitative conclusions – compared to a single research team that comes to a single, joint conclusion – to be more swayed by the findings, as it seems to reflect consensus between various approaches. After all, it’s a hallmark of rigorous research to triangulate various methodological approaches and provide converging evidence (Jick, 1979; Turner, Cardinal, & Burton, 2017), and to provide a “sensitivity analysis” or “robustness checks” to examine whether and how the findings change as a result of alternative analytic specifications (Muñoz & Young, 2018; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). As argued in Aczel et al. (2020, p. 562), “The main argument for the importance of performing robustness checks over reasonable variations in modelling choices is to increase confidence in the obtained results: ideally, results should be reasonably unaffected by a researcher’s idiosyncratic choice (…) when reasonable alternatives exist.” The same argument has been made for triangulation, which, according to Jick (1979, p. 608), “allows researchers to be more confident of their results.” We would similarly expect readers of multi-analyst reports that yield consistent findings to be more confident in the results. Although lay observers, in such cases, may be less certain that the precise aggregate parameter estimate is exactly correct, they should be more likely to positively update their beliefs about the reported phenomenon, and less likely to assume that the overall estimate stems from bias or error. Thus, when the results generated by independent analysts are largely consistent, we expect an increase in the sway of scientific findings. However, when laypeople observe several scientists independently come to differing qualitative conclusions, we expect the multi-analyst method to backfire; when results across many analysts vary widely and lack consensus in their qualitative conclusions (which, arguably, often reflects the reality of large-scale science collaborations), we expect a decrease in the sway of scientific findings.  
**[Insert Table 1 here]**  
Our pre-registered hypotheses (<https://osf.io/rpu98>) can be found in Table 2: we hypothesized that in the multi-consistent condition (compared to the single-analyst condition), lay consumers would have higher posterior beliefs, would find the results more credible, and would be less likely to believe the results stem from bias or error. For the multi-inconsistent condition, we hypothesized that lay consumers would have lower posterior beliefs, would find the results less credible, and would be more likely to believe the results stem from bias or error. In addition, we expected that the act of providing multiple (slightly to widely varying) parameter estimates would decrease confidence in the aggregate parameter estimate in both multi-analyst conditions.  
**[Insert Table 2 here]**

# 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 1,500 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%, 6%, and 8%, respectively) that average to 5% (SD = 2); 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% (SD = 8.25).  
Afterwards, participants rated (1) their posterior beliefs, (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) experimenter 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 2. We did not preregister any hypotheses for the last measure; the findings concerning the impact of experimental condition on ratings of experimenter 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

After two weeks of data collection, we recorded 2019 responses in Qualtrics. We excluded 120 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 328 participants who failed the attention check. This left us with a total sample of 1498 participants (499, 500, and 499 in the single-analyst, multi-consistent, and multi-inconsistent condition, respectively).  
Our main findings are displayed in Figure 1. Controlling for prior beliefs and comparing to the single-analyst condition, we found that (1) reported posterior 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 experimenter degrees of freedom were significantly greater in both the multi-consistent condition, , 95% CI , , , and the multi-inconsistent condition, , 95% CI , , .  
**[Insert Figure 1 here]**  
In line with our hypotheses, lay consumers of multi-analyst studies with inconsistent results (compared to single-analyst studies) have lower posterior beliefs, 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 lay consumers of multi-analyst studies with consistent results (compared to single-analyst studies) have higher posterior beliefs, find the results more credible, have less confidence in the average effect size estimate, and believe the results are less likely to stem from bias and error: instead, they report significantly lower posterior beliefs and are more likely to believe the results stem from error (we did not find significant effects on ratings of credibility, confidence, or bias). Figure 2 further clarifies the sway of multi-analyst vs. single-analyst studies, by displaying the distribution of prior and posterior beliefs across the three conditions.  
**[Insert Figure 2 here]**  
It is worth noting on the basis of Figure 2 and a post-hoc, paired *t*-test that, while multi-analyst studies with consistent results perform worse or no better than single-analyst studies on all measures, there is a significant, positive effect of the findings on posterior beliefs *within* the multi-consistent condition: i.e., beliefs in the research hypothesis are greater after reading the multi-consistent study results, , 95% CI , , . This finding clarifies that multi-analyst studies are not necessarily bad in absolute terms — however, when comparing to conventional, single-analyst scientific research, crowdsourced data analysis does not seem to provide an improvement in the sway and credibility of scientific research to lay consumers.

# 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 reliability of scientific research. The multi-analyst approach comes with many worthy uses, from demonstrating the arbitrariness and impact of individual analytic choices to acknowledging the inherent variability of results and averaging across idiosyncratic analytic choices to obtain more accurate parameter estimates. However, as with any real-world intervention, scientific reform can have unintended consequences. Here, we focus on the effects of crowdsourcing data analysis, and find that the multi-analyst approach may have an unintended consequence.  
While partly instituted with the goal of improving the credibility of scientific research (as shown in Table 1), lay consumers appear to resist the variability and lack of consensus that often comes with multi-analyst research. To our surprise, even when results generated by independent analysts are largely consistent, lay consumers are less likely to believe in the reported phenomenon and more likely to think that the findings stem from error and experimenter degrees of freedom.

## Acknowledgements

We thank Nicole Clare Kolmstetter for valuable assistance in the data collection. 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; Allen et al., 2021), *readr* (Version 1.4.0; Wickham & Hester, 2020), *rmarkdown* (Version 2.11; Xie, Allaire, & Grolemund, 2018; Xie, Dervieux, & Riederer, 2020), and *tinylabels* (Version 0.2.1; Barth, 2021).

## Disclosure statement

There are no relevant financial or non-financial competing interests to report.

## 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

Aczel, B., Hoekstra, R., Gelman, A., Wagenmakers, E.-J., Klugkist, I. G., Rouder, J. N., … van Ravenzwaaij, D. (2020). Discussion points for Bayesian inference. *Nature Human Behaviour*, *4*(6), 561–563. <https://doi.org/10.1038/s41562-019-0807-z>

Aczel, B., Szaszi, B., Nilsonne, G., Akker, O. van den, Albers, C., Assen, M. A. L. M. van, … Wagenmakers, E.-J. (2021). *Guidance for conducting and reporting multi-analyst studies*. MetaArXiv. <https://doi.org/10.31222/osf.io/5ecnh>

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

Arbon, R., Drax, K., Thurlby, N., Timpson, N. J., Northstone, K., Brown, K. R., … Munafo, M. (2019). *MAPS: Mapping the Analytical Paths of a Crowdsourced Data Analysis*. <https://doi.org/10.17605/OSF.IO/7VQ3B>

Auspurg, K., & Brüderl, J. (2021). Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the “Many Analysts, One Data Set” Project. *Socius*, *7*, 1–14. <https://doi.org/10.1177/23780231211024421>

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>

Breznau, N. (2021). I Saw You in the Crowd: Credibility, Reproducibility, and Meta-Utility. *PS: Political Science & Politics*, *54*(2), 309–313. <https://doi.org/10.1017/S1049096520000980>

Breznau, N., Rinke, E. M., Wuttke, A., Adem, M., Adriaans, J., Alvarez-Benjumea, A., … Nguyen, H. H. V. (2021). *Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty*. <https://doi.org/10.31222/osf.io/cd5j9>

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>

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>

Jick, T. D. (1979). Mixing Qualitative and Quantitative Methods: Triangulation in Action. *Administrative Science Quarterly*, *24*(4), 602–611. <https://doi.org/10.2307/2392366>

Muñoz, J., & Young, C. (2018). We Ran 9 Billion Regressions: Eliminating False Positives through Computational Model Robustness. *Sociological Methodology*, *48*(1), 1–33. <https://doi.org/10.1177/0081175018777988>

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>

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*.

Silberzahn, Raphael, & Uhlmann, E. L. (2015). Crowdsourced research: Many hands make tight work. *Nature*, *526*(7572), 189–191. <https://doi.org/10.1038/526189a>

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. <https://doi.org/10.1177/2515245917747646>

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. <https://doi.org/10.1177/1745691616658637>

Turner, S. F., Cardinal, L. B., & Burton, R. M. (2017). Research Design for Mixed Methods: A Triangulation-based Framework and Roadmap. *Organizational Research Methods*, *20*(2), 243–267. <https://doi.org/10.1177/1094428115610808>

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

## Table 1

## *Proposed credibility- and confidence-related benefits of crowdsourced data analysis*

|  |  |
| --- | --- |
| Citation | Description of credibility- or confidence-related benefits |
| (Aczel et al., 2021, pp. 3–4) | “When the results of independent data analyses converge, more confidence in the conclusions is warranted. When the results diverge, confidence appropriately falters, and scientists can examine the reasons for these discrepancies.” |
| (Arbon et al., 2019) | “For the public to have faith in the conclusions of scientists it is important that the methods they employ are robust and transparent. This study will examine robustness by recruiting teams of independent data analysts and looking at how they answer a controversial research question using the same data, effectively ‘crowd-sourcing’ the data analysis.” |
| (Auspurg & Brüderl, 2021, pp. 1, 10) | “Several researchers analyze the same research question with the same data (…) Science is credible if different researchers come up with a similar answer” |
| (Breznau et al., 2021, p. 3) | “Organized scientific knowledge production (…) should generate inter-researcher reliability, offering consumers of scientific findings assurance that they are not arbitrary flukes but that other researchers would generate similar findings given the same data.” |
| (Breznau, 2021, p. 311) | “crowdsourcing provides a new way to increase credibility for political and social research—in both sample populations and among the researchers themselves (…) It is hoped that these developments are tangible outcomes that increase public, private, and government views of social science.” |
| (Raphael Silberzahn & Uhlmann, 2015, pp. 190–191) | “the results are more credible (…) greater certainty comes from having an independent check.” |
| (R. Silberzahn et al., 2018, p. 352) | “Scientists can have comparatively more faith in a finding when there is less variability in analytic approaches taken to investigating the targeted phenomenon and in results obtained using different methods. (…) In such extreme cases of little to no convergence in results, the crowdsourcing process suggests that the scientific community should have no faith that the hypothesis is true” |
| (Uhlmann et al., 2019, p. 713) | “crowdsourced teams can conduct high-powered, precise studies and draw confident conclusions. (…) Crowdsourcing research is a part of a changing landscape of science that seeks to improve research reliability and advance the credibility of academic research” |

## Table 2

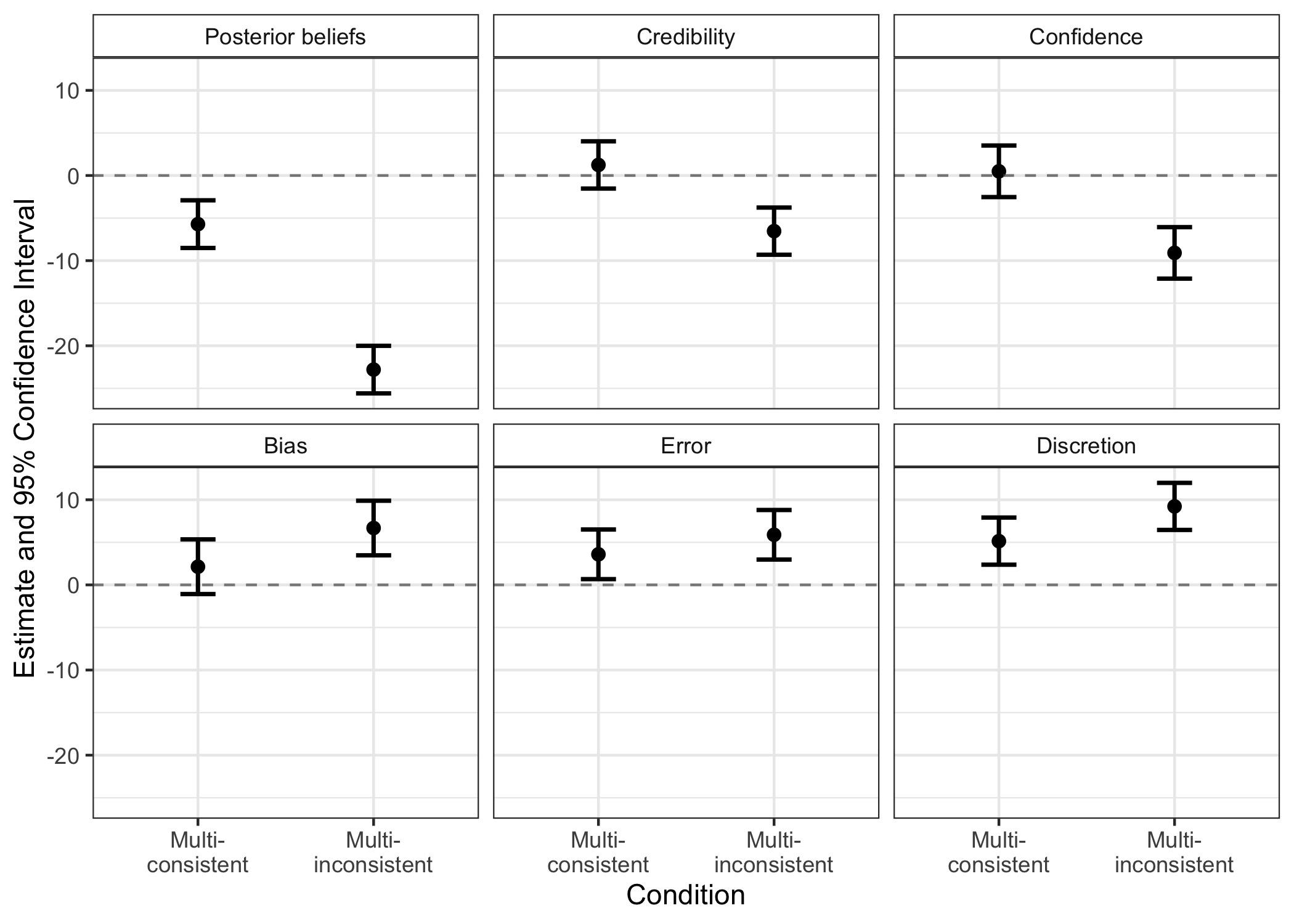
Predicted direction of effects for all dependent variables, compared to the single-analyst condition and controlling for prior beliefs

|  |  |  |
| --- | --- | --- |
| Measure | Many-analyst: Consistent | Many-analyst: Inconsistent |
| 1. Final beliefs | + | - |
| 2. Credibility | + | - |
| 3. Confidence | - | - |
| 4. Bias | - | + |
| 5. Error | - | + |
| 6. Discretion | No prediction | No prediction |

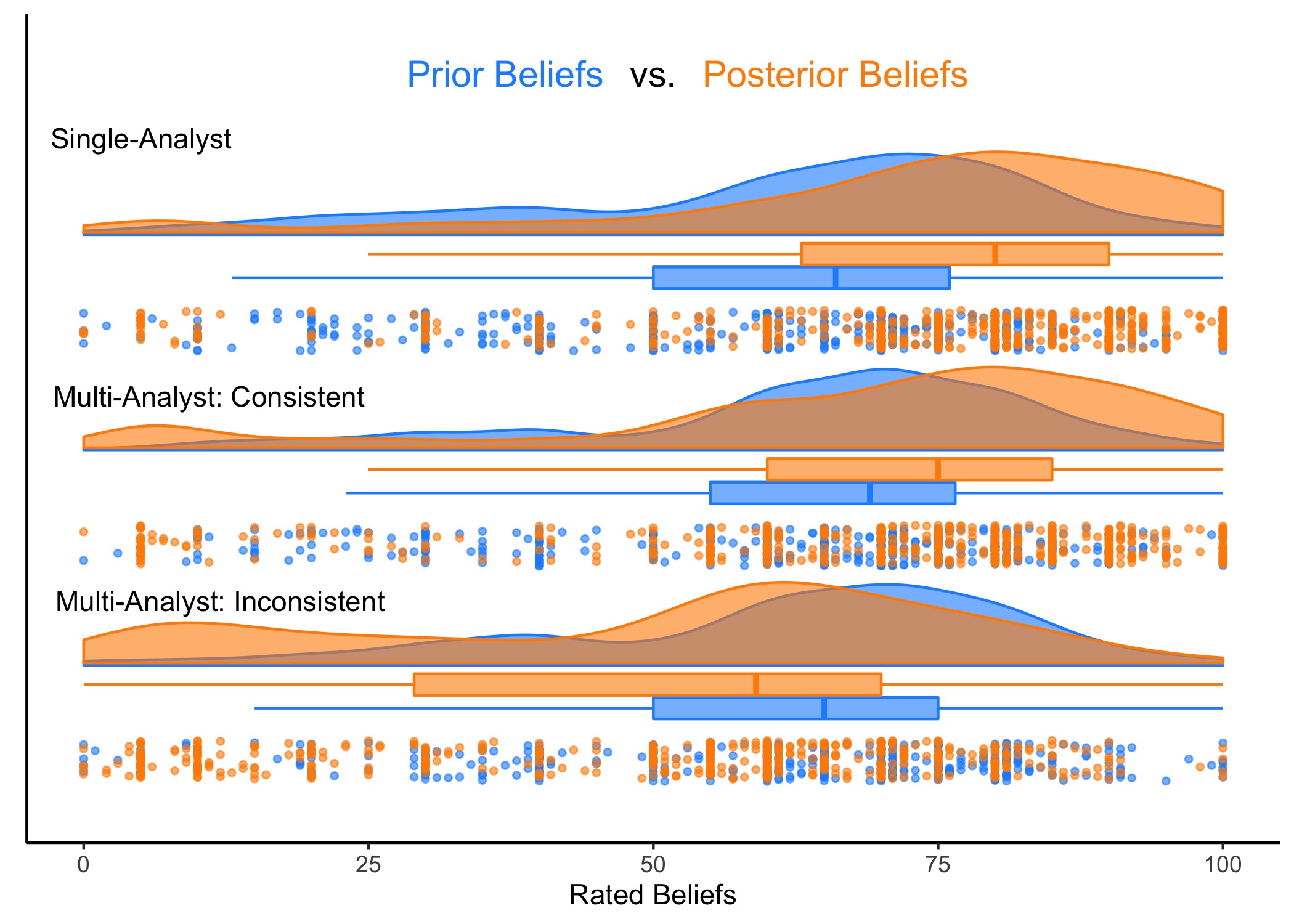
*Note*. Table 2 indicates the predicted direction of the effect for each of the five dependent variables, compared to the single-analyst condition and controlling for prior beliefs. For example, we hypothesized that, compared to a single-analyst study and controlling for prior beliefs, ratings of credibility would be greater in the multi-analyst: consistent condition and lower in the multi-analyst: inconsistent condition.

# Figures

## Figure 1

*Ratings of Bias, Confidence, Credibility, Discretion, Error, and Posterior Beliefs*  *Note.* Figure 1 displays coefficient estimates (and 95% confidence intervals) of posterior beliefs, credibility, confidence, bias, error, and discretion 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 posterior beliefs in the single-analyst, multi-consistent, and multi-inconsistent conditions*  
  *Note.*  Prior beliefs are displayed in blue; posterior beliefs are displayed in orange. The respective boxes display the lower quartiles, medians, and upper quartiles of prior and posterior beliefs by condition.