Lay Perceptions of Scientific Findings: Swayed by the Crowd?

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, confidence in an aggregate effect size estimate, and ratings of credibility, 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 estimate and report a parameter of interest. We compare the effects of providing a single, aggregate parameter estimate (the single estimate condition) vs. multiple parameter estimates that (a) vary slightly and are all positive, leading to the same qualitative conclusion (the consistent crowd condition) or (b) vary widely and are of both signs, leading to differing qualitative conclusions (the inconsistent crowd 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:* 2,222

Lay Perceptions of Scientific Findings: Swayed by the Crowd?

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: “the organization of scientific research in open and collaborative projects” (Franzoni & Sauermann, 2014, p. 1). We focus on crowdsourced data analysis, also known as the ‘many analysts’ or ‘multi-analyst’ approach: a crowd of scientists independently analyzes the same dataset to estimate and report on a parameter of interest.  
According to some of its proponents, crowdsourced data analysis should increase the credibility of scientific findings to lay consumers. In the preregistration of an ongoing many-analysts study, Arbon and colleagues (2019) justify their use of the novel approach by referring to the need for “the public to have faith in the conclusions of scientists.” In a preprint of a conducted multi-analyst study, Breznau and colleagues (2021, p. 3) argue that independent data analyses with converging findings should offer “consumers of scientific findings assurance that they are not arbitrary flukes.” In a commentary on the many analysts approach, Breznau (2021, p. 311), offers the hopeful view that crowdsourcing is a new way to “increase public, private, and government views of social science” and to “increase credibility for political and social research—in both sample populations and among the researchers themselves.” Does crowd science meet this promise – to improve lay perceptions of scientific findings – 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 compare the effects of providing lay consumers with a single, aggregate parameter estimate (the single estimate condition) vs. multiple parameter estimates that (a) vary slightly and are all positive, leading to the same qualitative conclusion (the consistent crowd condition) or (b) vary widely and are of both signs, leading to differing qualitative conclusions (the inconsistent crowd condition).  
In line with social norms theory (Miller & Prentice, 2016), we expect that observing consensus among a crowd (the consistent crowd condition) will – compared to the conclusion of a single scientist (the single estimate condition) – increase conformity in opinion. Drawing from work on intuitive statistics (Gigerenzer & Murray, 2015), we also expect laypeople to intuitively accord to the logic of “the wisdom of crowds”: the ability of an aggregate of estimates (rather than a single estimate) to reduce noise stemming from individual bias or error (Schweinsberg et al., 2021).  
In contrast, when crowd estimates show low consensus and high variance (the inconsistent crowd condition), we predict that observers will be less swayed and more likely to attribute the findings to bias and error. In addition, due to the difficulty of lay reasoning about variation (Garfield & Ben-Zvi, 2004), we predict an aversion to variability: i.e., we expect that observing variable estimates will decrease lay confidence in the precise average parameter estimate in both crowd conditions. In sum, 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 1: we hypothesized that in the consistent crowd condition (compared to the single estimate condition, and controlling for prior beliefs), 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 inconsistent crowd condition, we hypothesized that (compared to the single estimate condition, and controlling for prior beliefs) 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 crowd conditions.  
In sum, our preregistered hypotheses are as follows: when laypeople observe multiple consistent (inconsistent) estimates from a crowd of independent scientists, we expect – compared to a single estimate and controlling for prior beliefs – higher (lower) posterior beliefs and credibility of the results, lower confidence in the precise average parameter estimate, and lower (greater) 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 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 estimate condition, a single team of six researchers reports a 5% increase in well-being among religious people; in the consistent crowd condition, six independent researchers report six consistent estimates (2%, 4%, 5%, 5%, 6%, and 8%, respectively) that average to 5% ( = 2%); and in the inconsistent crowd condition, six independent researchers report six inconsistent estimates (-6%, -2%, 5%, 5%, 12%, and 16%, respectively) that average to 5% ( = 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) researcher 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 estimate 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 researcher 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 estimate vs. the consistent crowd condition, and one comparison of the single estimate vs. the inconsistent crowd 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 estimate, consistent crowd, and inconsistent crowd condition, respectively).  
Figure 1 displays the main findings of our linear regression models, which compare each condition to the single estimate condition and control for prior beliefs. For the inconsistent crowd condition, we found significantly lower (a) posterior beliefs, , 95% CI , ; (b) ratings of credibility, , 95% CI , ; and (c) confidence in the effect size estimate, , 95% CI , ; and significantly higher ratings of (d) bias, , 95% CI , ; (e) error, , 95% CI , ; and (f) discretion, , 95% CI , .  
For the consistent crowd condition, we found significantly lower (a) posterior beliefs, , 95% CI , ; significantly higher ratings of (b) error, , 95% CI , ; and (c) discretion, , 95% CI , ; and no significant effects on (d) ratings of credibility, , 95% CI , ; (e) confidence in the effect size estimate, , 95% CI , ; or (f) ratings of bias, , 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 positive difference between prior and posterior beliefs within the consistent crowd condition: i.e., beliefs in the research hypothesis are greater after consuming consistent crowd estimates, , 95% CI , . This finding clarifies that a consistent crowd may sway observers — however, compared to a conventional, non-crowdsourced estimate, crowd estimates do not seem to improve lay perceptions of scientific findings.

# Discussion

Every day, important scientific findings are rejected at large. From man-made climate change to the safety and efficacy of Covid-19 vaccinations, science skepticism has run rampant among lay consumers in modern society (Hornsey & Fielding, 2017). To increase public faith in science, some have proposed the use of crowd science (Silberzahn et al., 2018; Uhlmann et al., 2019). In Table 2, we provide an overview of the proposed confidence- and credibility-related benefits of the multi-analyst approach. The logic behind the crowd science approach accords with that of analytic practices that are considered to be hallmarks of rigorous research: e.g., 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.” The same argument has been made for triangulation, which, according to Jick (1979, p. 608), “allows researchers to be more confident of their results.” Drawing from theories on social norms and numerical cognition, we similarly expected lay consumers of multi-analyst studies with consistent findings to be more confident in the results.  
**[Insert Table 2 here]**  
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. 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, 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 highly consistent and lead to the same qualitative conclusion, we find no improvement in lay perceptions of scientific findings. Instead, consumers of consistent, slightly variable, crowd estimates are less likely to believe in the reported phenomenon and more likely to think that the findings stem from error and researcher degrees of freedom. In the future, it is important for crowd scientists to consider how to tackle science skepticism and effectively communicate variable estimates.

## Acknowledgements

We thank Nicole Clare Kolmstetter for valuable assistance in the data collection. This manuscript was created using R (Version 4.1.2; R Core Team, 2021) and the R-packages *broom* (Version 0.7.10; Robinson, Hayes, & Couch, 2021), *dplyr* (Version 1.0.7; Wickham, François, Henry, & Müller, 2021), *ggplot2* (Version 3.3.5; Wickham, 2016), *glue* (Version 1.5.1; Hester, 2020), *here* (Version 1.0.1; Müller, 2020), *icons* (Version 0.2.0; O’Hara-Wild, 2021), *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 2.0.2; Wickham & Hester, 2020), *rmarkdown* (Version 2.11; Xie, Allaire, & Grolemund, 2018; Xie, Dervieux, & Riederer, 2020), *rsvg* (Version 2.1.2; Ooms, 2021), 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*. MetaArXiv. <https://doi.org/10.31222/osf.io/cd5j9>

Franzoni, C., & Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. *Research Policy*, *43*(1), 1–20. <https://doi.org/10.1016/j.respol.2013.07.005>

Garfield, J., & Ben-Zvi, D. (2004). Research on Statistical Literacy, Reasoning, and Thinking: Issues, Challenges, and Implications. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 397–409). Dordrecht: Springer Netherlands. <https://doi.org/10.1007/1-4020-2278-6_17>

Gigerenzer, G., & Murray, D. J. (2015). *Cognition as Intuitive Statistics*. Psychology Press.

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>

Miller, D. T., & Prentice, D. A. (2016). Changing Norms to Change Behavior. *Annual Review of Psychology*, *67*(1), 339–361. <https://doi.org/10.1146/annurev-psych-010814-015013>

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

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>

O’Hara-Wild, M. (2021). *Icons: SVG icons for r documents and apps*.

Ooms, J. (2021). *Rsvg: Render SVG images into PDF, PNG, PostScript, or bitmap arrays*. Retrieved from <https://CRAN.R-project.org/package=rsvg>

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*.

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C. M., van Assen, M. A. L. M., … et al. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes*, *165*, 228–249. <https://doi.org/10.1016/j.obhdp.2021.02.003>

Silberzahn, R., & 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

*Predicted direction of effects compared to the single estimate condition*

| Measure | Consistent crowd | Inconsistent crowd |
| --- | --- | --- |
| 1. Posterior beliefs |  |  |
| 2. Credibility of the results |  |  |
| 3. Confidence in precise estimate |  |  |
| 4. Researcher bias |  |  |
| 5. Researcher error |  |  |
| 6. Researcher discretion | No prediction | No prediction |

*Note*. Table 1 indicates the predicted direction of the effect for all dependent variables, compared to the single estimate condition and controlling for prior beliefs (a green plus stands for a positive prediction, and a red minus stands for a negative prediction). For example, we hypothesized that, compared to a single-analyst study and controlling for prior beliefs, ratings of credibility would be greater in the consistent crowd condition and lower in the inconsistent crowd condition.

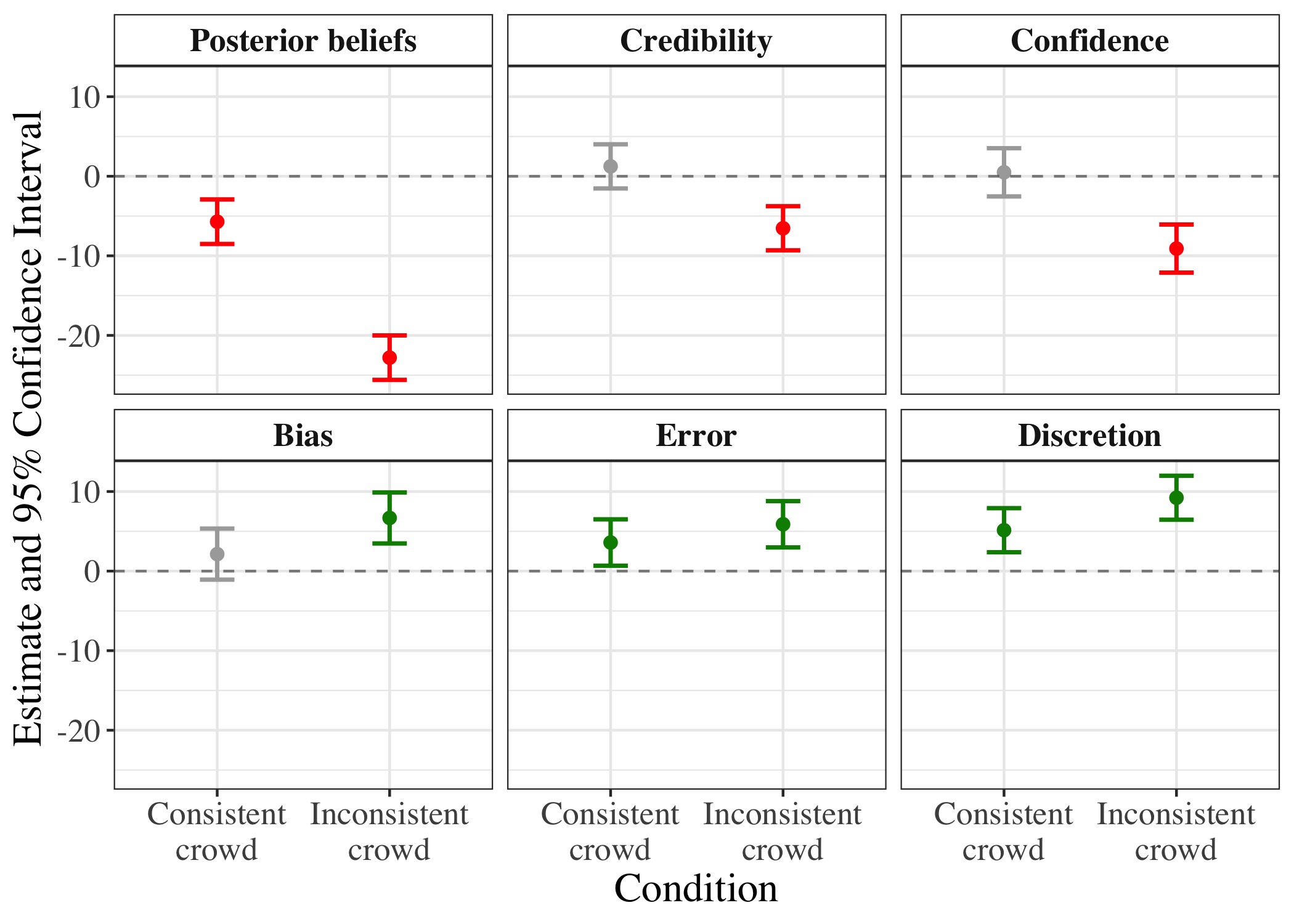
## Table 2

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

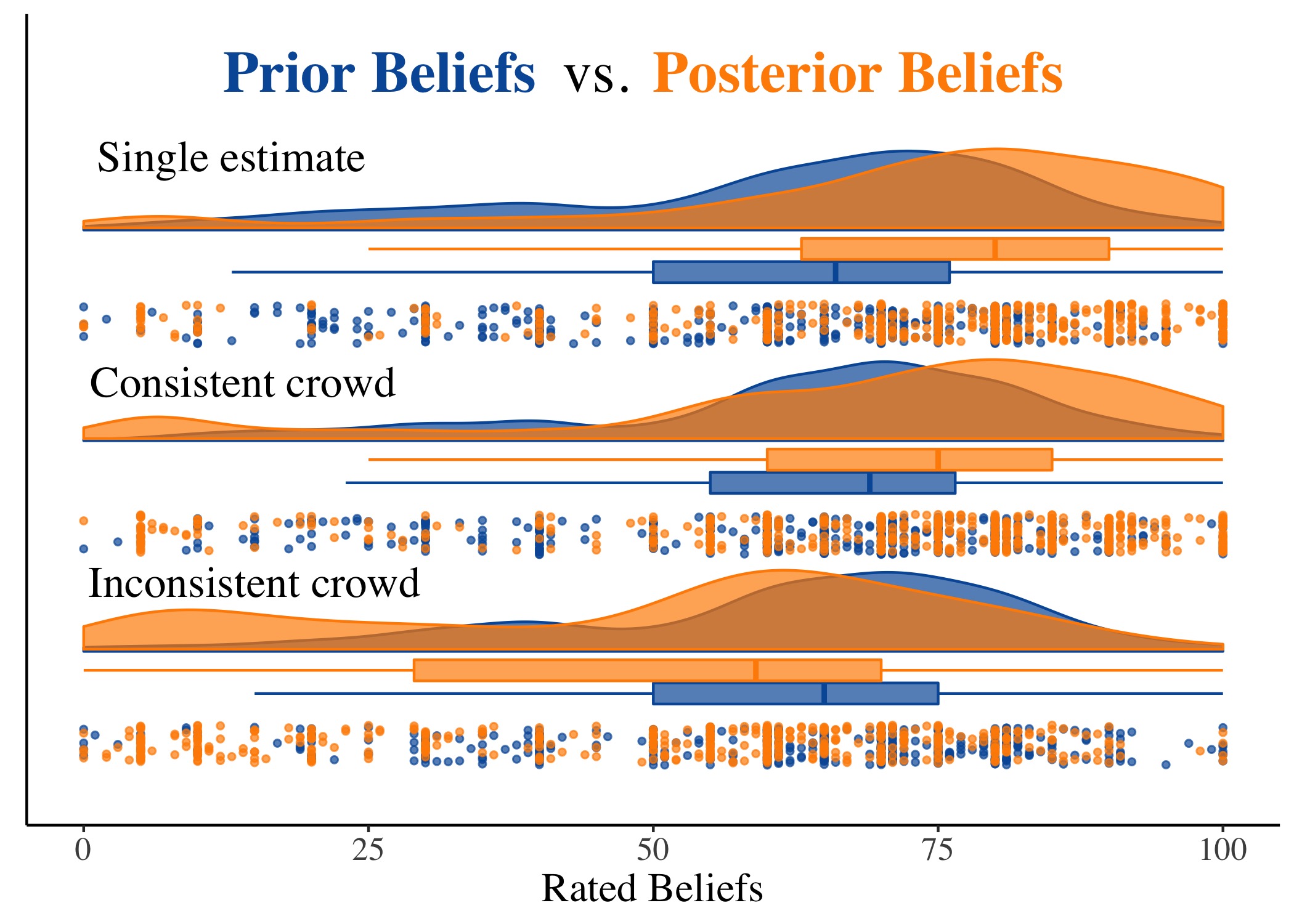
| Citation | Description of credibility- or confidence-related benefits |
| --- | --- |
| (Aczel et al., 2021, pp. 3–4) | “When the results of independent data analyses converge, more confidence in the conclusions is warranted. When the results diverge, confidence appropriately falters, and scientists 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.” |
| (Silberzahn & Uhlmann, 2015, pp. 190–191) | “the results are more credible (…) greater certainty comes from having an independent check.” |
| (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” |

# Figures

## Figure 1

*Estimates of differences with the single estimate condition*  *Note.* We regressed all outcomes on prior beliefs and condition (with the single estimate condition as the reference category). Figure 1 displays coefficient estimates (and 95% confidence intervals) of posterior beliefs, credibility, confidence, bias, error, and discretion in the two crowd conditions – the consistent crowd and the inconsistent crowd – compared to the single-analyst condition and controlling for prior beliefs. Green/red/gray error bars indicate positive/negative/insignificant findings, respectively.

## Figure 2

*Individual data points, quartiles, and distributions of prior and posterior beliefs in the single estimate, consistent crowd, and inconsistent crowd 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.