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

Insert first author name here1 & Insert second author name here1

1 Insert affiliation here

Abstract

Crowd science is on the rise in behavioral science and aims to increase the credibility of scientific research. Whether it does, however, remains untested. We run an experiment to study whether crowd science improves lay perceptions of scientific findings. We focus on crowdsourced data analysis: a crowd of scientists independently analyzes the same dataset to estimate and report on a parameter of interest.  
We compare the effects of providing lay consumers (*N* = 1,498) with a single non-crowdsourced parameter estimate (the single-analyst condition) vs. multiple crowdsourced parameter estimates that (a) vary slightly and are all positive, leading to the same qualitative conclusion (the “multi-consistent” condition) or (b) vary widely and are of both signs, leading to differing qualitative conclusions (the “multi-inconsistent” condition). In the single-analyst condition, a single team of six researchers reports a 5% estimate; in the multi-consistent condition, six independent researchers report six low variance, high consensus estimates (2%, 4%, 5%, 5%, 6%, 8%; ); in the multi-inconsistent condition, six independent researchers report six high variance, low consensus estimates (-6%, -2%, 5%, 5%, 12%, 16%; ).  
According to our preregistered hypotheses and in line with social norm theory, when laypeople observe multiple consistent (inconsistent) estimates from independent scientists rather than a single estimate, we expect an increase (decrease) in posterior beliefs and credibility of the results, and a decrease (increase) in ratings of bias and error. In addition, we expect that observing differing estimates decreases confidence in the precise average estimate in both multi-estimate conditions.  
In line with our hypotheses, we find that lay consumers of multiple, inconsistent (high variance, low consensus) crowdsourced estimates have lower posterior beliefs, , 95% CI , ; find the results less credible, , 95% CI , ; have less confidence in the average effect size estimate, , 95% CI , ; and believe the results are more likely to stem from bias, , 95% CI , , and error, , 95% CI , , than consumers of a single, non-crowdsourced estimate. Contrary to our hypotheses, we do not find that consuming multiple, consistent (low variance, high consensus) crowdsourced estimates improves lay perceptions of scientific findings: instead, to our surprise, it leads to lower posterior beliefs, , 95% CI , , and higher ratings of error, , 95% CI , , than a single, non-crowdsourced estimate.

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

*Word count:* Abstract: 383, Project summary: x

Lay Perceptions of 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: “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 et al. (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 et al. (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 parameter estimate (the single-analyst condition) vs. multiple crowdsourced parameter estimates that (a) vary slightly and are all positive, leading to the same qualitative conclusion (the “multi-consistent” condition) or (b) vary widely and are of both signs, leading to differing qualitative conclusions (the “multi-inconsistent” condition).  
In line with social norm theory (Miller & Prentice, 2016), we expect an individual who observes a crowd of scientists reach identical (differing) qualitative conclusions to be more (less) likely to conform in opinion, and be more (less) swayed by the scientific findings, than an individual who observes the conclusion of a single scientist. However, regardless of the presence or absence of consensus in conclusions, we expect that observing (slightly to widely) varying estimates decreases lay confidence in the precise average estimate. 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, non-crowdsourced estimate – higher (lower) posterior beliefs and credibility of the results, lower (lower) confidence in the precise average parameter estimate, and lower (greater) ratings of bias and error.

# Methods

All relevant materials – including the preregistration, survey materials, data, and code – are openly available on the OSF at [insert link]. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. We sampled UK- and US-based participants from the Prolific participant pool (native English speakers, over 18 years old). We excluded participants who failed an attention check or attempted to take the survey more than once. Our target sample size was 1,500 participants after exclusions. 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) 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-analyst condition as the reference category) and prior beliefs as a covariate. All hypotheses, statistical models, and code were preregistered. 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-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).  
Figure 1 displays the main findings of our linear regression models, which compare each condition to the single-analyst condition and control for prior beliefs. For the multi-inconsistent 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 (d) ratings of bias, , 95% CI , ; (e) error, , 95% CI , ; and (f) discretion, , 95% CI , .  
For the multi-consistent condition, we found significantly lower (a) posterior beliefs, , 95% CI , ; significantly higher (b) ratings of 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 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 has 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, 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 researcher degrees of freedom.

# References

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>

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>

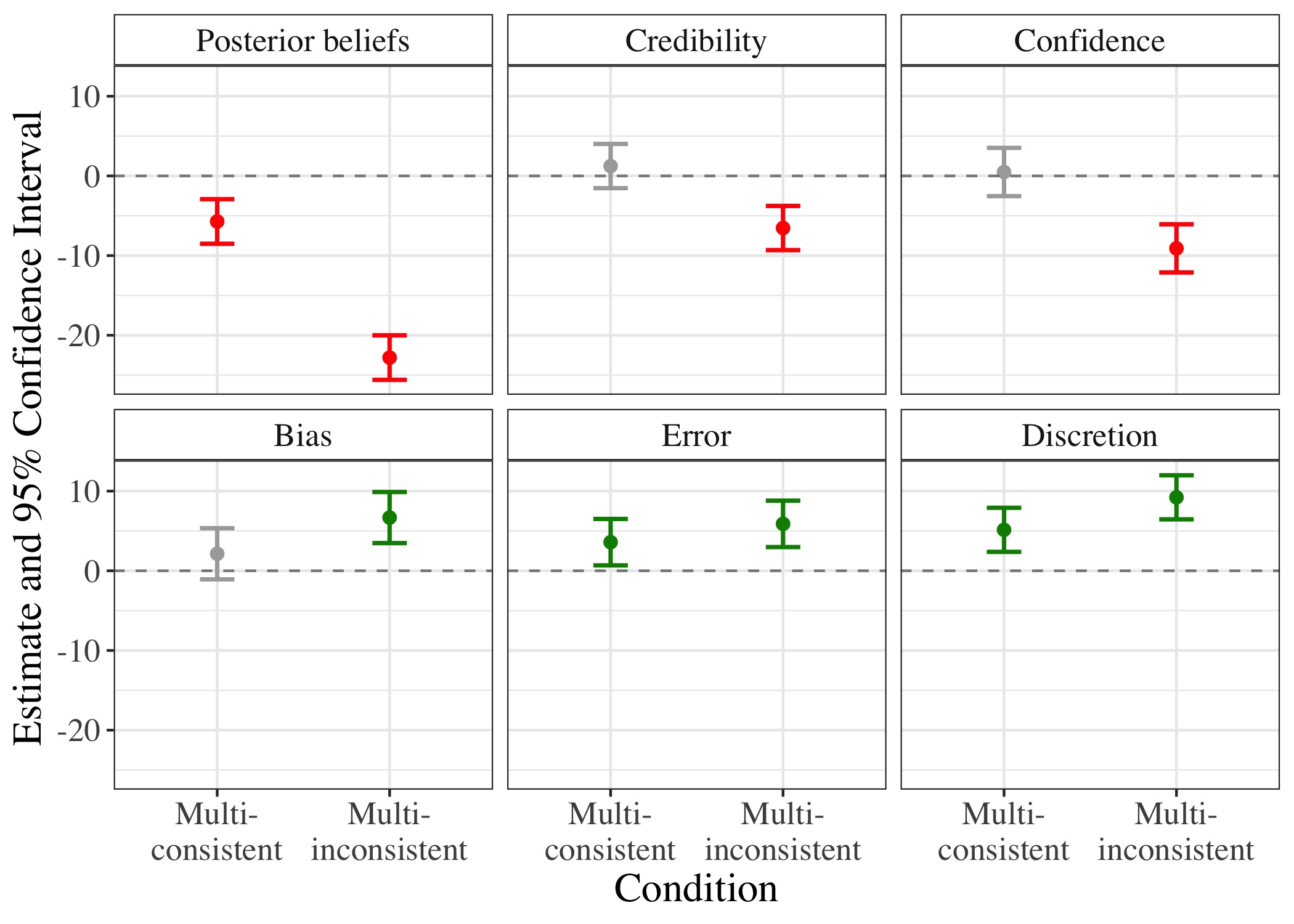
Hornsey, M. J., & Fielding, K. S. (2017). Attitude roots and Jiu Jitsu persuasion: Understanding and overcoming the motivated rejection of science. *The American Psychologist*, *72*(5), 459–473. <https://doi.org/10.1037/a0040437>

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>

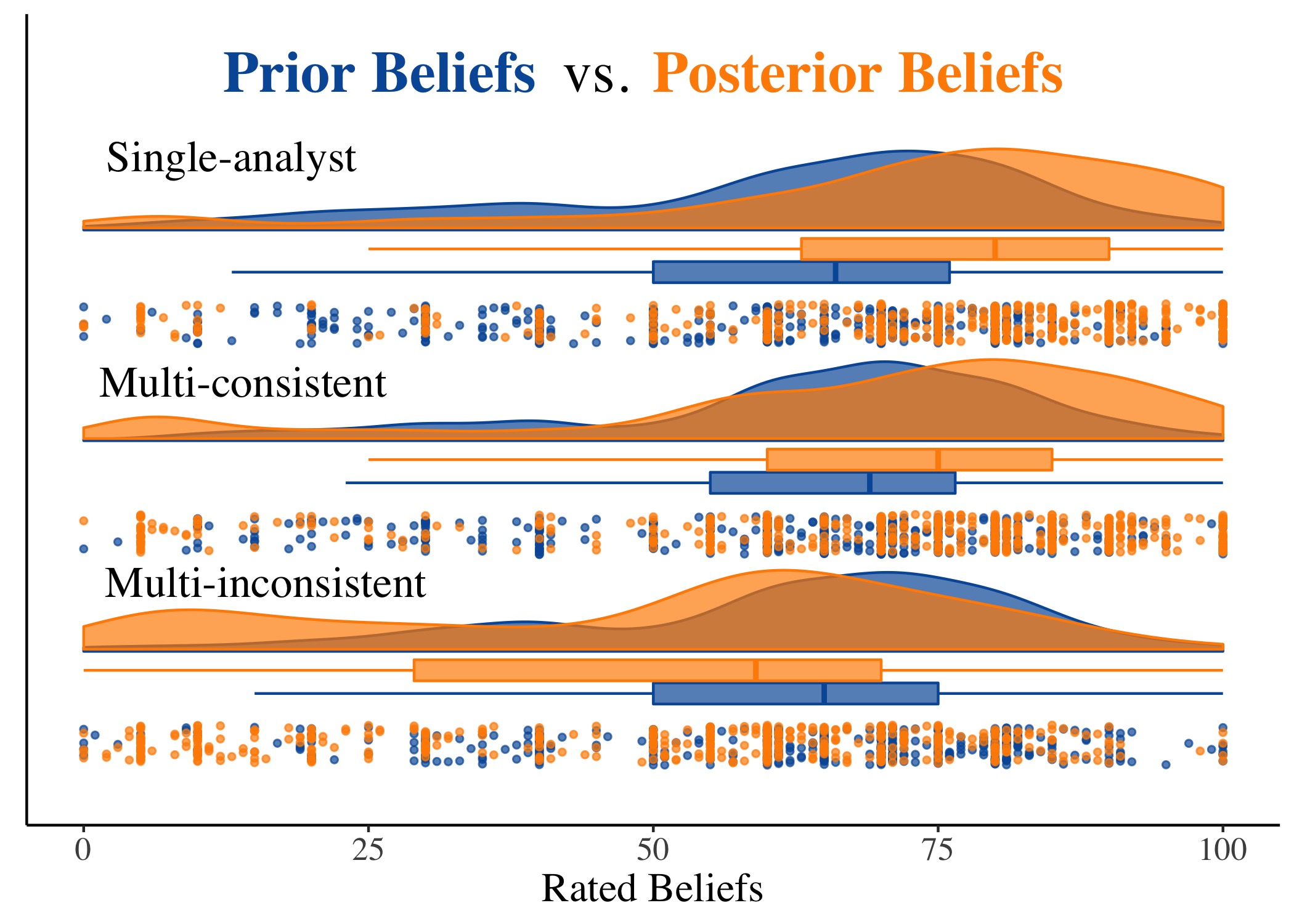
Pashler, H., & Wagenmakers, E. (2012). Editors’ Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. <https://doi.org/10.1177/1745691612465253>

# Figures

## Figure 1

*Estimates relative to the single-analyst condition*  *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). 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-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.