SM2: Consulting field experts to identify potential quantitative indicators of uncertainty

Peder M. Isager

9/17/2021

To better understand what information is important for assessing overall uncertainty about findings from fMRI research, we constructed a survey to probe experts in fMRI research about which information they use to assess the quality and quantity of evidence for fMRI findings in their field. The purpose of the survey was twofold. First, we wanted an opportunity to discover quantitative indicators of uncertainty we had not previously considered, and that might be feasible to code in our data. Second, we wanted to compare the reported importance of sample size for evaluating uncertainty in comparison with other information researchers might also be using.

Methods

Pilot data collection was carried out on a convenience sample of colleagues of the first (Peder) and second (Anna) author. 13 researchers responded to the survey. All participants were researchers with, or in the process of completing, a PhD, who had experience with collecting and analyzing fMRI data. The purpose of this sampling restriction was to ensure that all participants had sufficient prior knowledge of fMRI methodology to give informative answers to the survey items they were presented with.

The survey was created in Qualtrics (https://www.qualtrics.com/). The survey and all data collected are available on OSF (https://osf.io/f7zdq/). The survey contained open-ended items encouraging researchers to list whatever information they considered important for assessing evidence. The survey also contained a number of closed-ended questions asking researchers to rate (on a visual analogue scale from 1:100 with 1 being the least important and 100 being the most important) and rank-order the importance of the following factors for judging the quality and quantity of evidence in support of a finding (this list of factors were generated by the authors after internal discussion about which factors could plausibly be used to evaluate uncertainty about social fMRI research):

1. The total sample size collected for the study.
2. The percentage of participants that were excluded (after they met the inclusion criteria for participating in MRI research).
3. The statistical power of the study to detect effect sizes of interest.
4. The size of the effect (e.g. Cohen's d for condition differences, Pearson's r for brain-behavior correlations, or percentage signal change for raw BOLD signal differences).
5. Cluster extent of relevant cluster(s).
6. The p-value for relevant cluster(s).
7. Whether the finding is a main effect or an interaction.
8. How participants were assigned to conditions, if relevant (e.g., randomly, single/double blind, etc.).
9. In cases where you know of a replication study, the result(s) of a replication study,
10. In cases where you know of a replication study, whether the replication is a close (direct) or conceptual replication.

11. In cases where you know of a replication study, whether the replication is conducted by an independent team or not.
12. Whether the finding is based on within-subjects measurements or between-subjects measurements.
13. Peak Z-value for relevant clusters.
14. Open access to the underlying empirical data that were analyzed.
15. Whether the study has been preregistered.
16. Whether there are statistical errors in the results reported (e.g., the degrees of freedom do not correspond to the other reported statistics, the total sample size does not equal the sum of the group sample sizes, etc.).
17. Whether the finding has a strong connection with theory.
18. Whether the finding is predicted a priori or discovered during data exploration.
19. How participants were sampled from the population (e.g., stratified random sampling, snowball sampling, convenience sampling, etc.).
20. Whether the finding is unexpected (e.g., counterintuitive), or in line with what we already know.

For each factor, we also asked for open-ended comments to better understand how the information was being used by researchers to assess evidence. For example, after asking researchers to rate the importance of "the percentage of participants that were excluded," we also asked participants to "indicate in what way you believe this information is related to the quality and quantity of evidence in support of a finding." We used the participants' responses on the items related to "the total sample size collected for the study" as a preliminary validation of whether sample size relates to uncertainty in the way assumed by Isager et al. (2020).

Results

The open responses by participants did not reveal novel quantitative indicators of uncertainty that we had not already considered, and that would be feasible to collect for all studies in our data.

There seemed to be broad agreement among experts that sample size is important for evaluating the quality and quantity of evidence for a typical fMRI finding. Several experts freely offered sample size as a piece of information they would be evaluating when assessing the credibility of a finding (before seeing our list of potentially important factors). Sample size also received the second highest median rating out of all factors (see table SM2-1 and figure SM2-1, and the highest average rank-order out of all factors (only "the results of a replication study" received an equally high rank. See table SM2-1 and figure SM2-2). In addition, statistical power, partially a function of sample size, was consistently highly rated and ranked by experts, and one expert explicitly pointed to the relationship between sample size and power in their comments ("Sample size is the easiest way to increase statistical power"). Finally, when asked specifically about the importance of sample size, there seemed to be broad agreement that a higher sample size generally entails higher credibility, in line with the assumptions of Isager, Veer, and Lakens (2021). However, two experts described feeling less confident about findings supported by a very high sample size, due to the elevated risks of overinterpreting trivially small and meaningless effects (a problem often referred to as "the crud factor," Meehl 1990; Orben and Lakens 2020). Besides sample size (and statistical power) participants seemed to consistently agree on the importance of a few other factors (see table SM2-1, figure SM2-1 and SM2-2).

There seemed to be broad agreement among experts that sample size is important for evaluating the quality and quantity of evidence for a typical fMRI finding. Several experts freely offered sample size as a piece of information they would be evaluating when assessing the credibility of a finding (before seeing our list of potentially important factors). Sample size also received the second highest median rating out of all factors (see table SM2 and figure SM2-1, and the highest average rank-order out of all factors (only "the results of a replication study" received an equally high rank. See table SM2 and figure SM2-2). In addition, statistical power, partially a function of sample size, was consistently highly rated and ranked by experts, and one expert explicitly pointed to the relationship between sample size and power in their comments ("Sample size is the easiest way to increase statistical power"). Finally, when asked specifically about the importance of

sample size, there seemed to be broad agreement that a higher sample size generally entails higher credibility, in line with the assumptions of Isager, Veer, and Lakens (2021). However, two experts described feeling less confident about findings supported by a very high sample size, due to the elevated risks of overinterpreting trivially small and meaningless effects (a problem often referred to as "the crud factor," Meehl 1990; Orben and Lakens 2020). Besides sample size (and statistical power) participants seemed to consistently agree on the importance of a few other factors (see table SM2, figure SM2-1 and SM2-2).

Overall, we cautiously interpret these results as preliminary validation of correspondence between the rationale of Isager et al. (2020) and how experts actually use sample size when evaluating uncertainty. However, we stress that the low sample size and exploratory nature of this pilot calls for replication before any firm conclusions can be drawn.

Table SM2. Median rating and rank for all factors asked about in the pilot survey.

Factor	N ratings	Median rating	N rankings	Median rank
effect predicted or exploratory	9	88	10	7
statistical errors	9	86	10	9.5
sample size	11	84	10	4
replication result	10	84	10	4
open data available	9	83	10	8
strongly connected to theory	10	81	10	4.5
statistical power	10	78.5	10	5.5
replication close or not	10	78	10	NA
preregistered	11	77	10	9.5
condition assignment	11	75	10	13
replication independent or not	10	74.5	10	NA
within or between design	9	71	10	13
effect size	9	67	10	5
cluster p-value	9	60	10	13
effect unexpected	9	59	10	10
cluster extent	10	51	10	12
participants excluded	10	45	10	15
cluster peak Z-value	9	37	10	11.5
participant sampling	10	24	10	15.5
main effect or interaction	9	17	10	12.5

References

Isager, Peder M., Anna van 't Veer, and Daniel Lakens. 2021. "Replication Value as a Function of Citation Impact and Sample Size." MetaArXiv. https://doi.org/10.31222/osf.io/knjea.

Meehl, Paul E. 1990. "Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles That Warrant It." *Psychological Inquiry* 1 (2): 108–41. https://doi.org/10.1207/s15327965pli0102_1.

Orben, Amy, and Daniël Lakens. 2020. "Crud (Re)Defined." *Advances in Methods and Practices in Psychological Science* 3 (2): 238–47. https://doi.org/10.1177/2515245920917961.

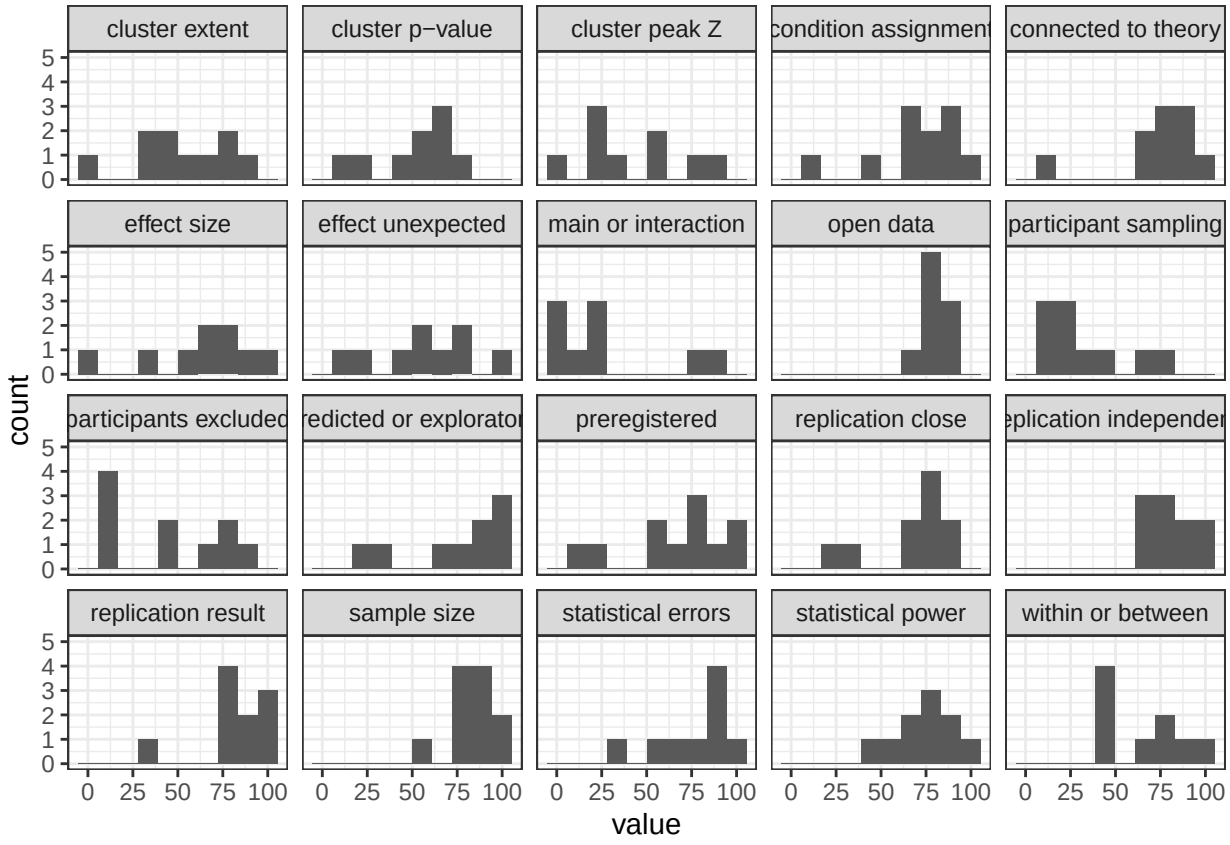

Figure SM2-1. Histograms of ratings for each of the 20 factors presented to participants. All factors were rated on a visual analogue scale from 1:100 with 1 being the least important and 100 being the most important.

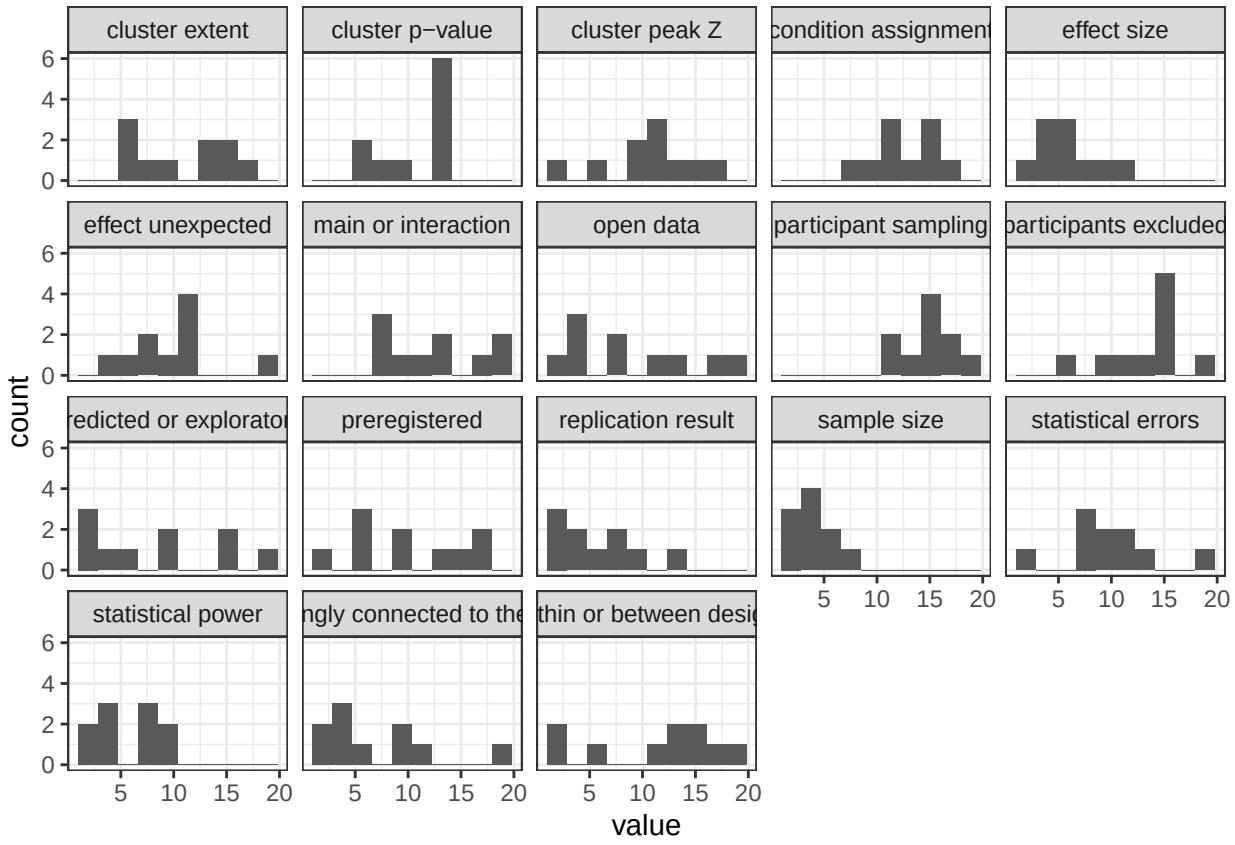

Figure SM2-2. Histograms of relative rank-order ratings for the 18 factors that participants were asked to rank ("replication close or not" and "replication independent or not" were not included for rank-ordering). Participants assigns one value of rank from 1-18 to all factors, where rank 1 indicates the most important factor, and rank 18 indicates the least important factor.