ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Sarafoglou A *et al*. 2024 Subjective evidence evaluation survey for many-analysts studies. *R. Soc. Open Sci.* **11**: 240125.
https://doi.org/10.1098/rsos.240125

Received: 1 January 2024
Accepted: 22 April 2024

Subject Category:
Science, society and policy

Subject Areas:
psychology/behaviour

Keywords:
open science, team science, scientific transparency, metascience, crowdsourcing analysis

Author for correspondence:
Alexandra Sarafoglou
e-mail: alexandra.sarafoglou@gmail.com

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.7360072.

THE ROYAL SOCIETY PUBLISHING

Subjective evidence evaluation survey for many-analysts studies

Alexandra Sarafoglou[1], Suzanne Hoogeveen[2], Don van den Bergh[1], Balazs Aczel[3], Casper J. Albers[4], Tim Althoff[5], Rotem Botvinik-Nezer[6,29], Niko A. Busch[7], Andrea M. Cataldo[8,30], Berna Devezer[9], Noah N. N. van Dongen[1], Anna Dreber[10,13], Eiko I. Fried[11], Rink Hoekstra[33], Sabine Hoffman[12], Felix Holzmeister[13], Jürgen Huber[13], Nick Huntington-Klein[14], John Ioannidis[15], Magnus Johannesson[10], Michael Kirchler[13], Eric Loken[16], Jan-Francois Mangin[17,31], Dora Matzke[1], Albert J. Menkveld[18], Gustav Nilsonne[19], Don van Ravenzwaaij[4], Martin Schweinsberg[20], Hannah Schulz-Kuempel[21,32], David R. Shanks[22], Daniel J. Simons[23], Barbara A. Spellman[24], Andrea H. Stoevenbelt[33], Barnabas Szaszi[3], Darinka Trübutschek[25], Francis Tuerlinckx[26], Eric L. Uhlmann[27], Wolf Vanpaemel[26], Jelte Wicherts[28] and Eric-Jan Wagenmakers[1]

[1]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands
[2]Utrecht University, Utrecht, The Netherlands
[3]Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary
[4]Heymans Institute for Psychological Research, University of Groningen, Groningen, The Netherlands
[5]Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA
[6]Hebrew University of Jerusalem, Jerusalem, Israel
[7]Institute for Psychology, University of Münster, Münster, Germany

[8]Center for Depression, Anxiety and Stress Research, McLean Hospital, Belmont, MA, USA
[9]Department of Business, University of Idaho, Moscow, ID, USA
[10]Stockholm School of Economics, Stockholm, Sweden
[11]Department of Psychology, Leiden University, Leiden, The Netherlands
[12]Department of Statistics, Ludwig-Maximilians-Universität München, Munchen, Bayern, Germany
[13]University of Innsbruck, Innsbruck, Tirol, Austria
[14]Seattle University, Seattle, WA, USA
[15]Meta-Research Innovation Center at Stanford (METRICS) and Departments of Medicine, of Epidemiology and of Population Health, of Biomedical Data Science, and of Statistics, Stanford University, Stanford, CA, USA
[16]University of Conneticut, Storrs, CT, USA
[17]University Paris-Saclay, Gif-sur-Yvette, France
[18]Vrije Universiteit Amsterdam, Amsterdam, Noord-Holland, The Netherlands
[19]Karolinska Institutet, Solna, Sweden
[20]ESMT Berlin, Berlin, Germany
[21]Department of Statistics and The Institute for Medical Information Processing, Biometry, and Epidemiology, LMU Munich, Munchen, Bayern, Germany
[22]Division of Psychology and Language Sciences, University College London, 26 Bedford Way, London WC1H 0AP, UK
[23]University of Illinois—Urbana-Champaign, Urbana, IL, USA
[24]School of Law, University of Virginia, 580 Massie Road, Charlottesville, VA, USA
[25]Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany
[26]University of Leuven, Leuven, Belgium
[27]INSEAD, Fontainebleau, Île-de-France, France
[28]Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands
[29]Dartmouth College, Hanover, NH, USA
[30]Department of Psychiatry, Harvard Medical School, Boston, MA, USA
[31]Neurospin CEA, Gif-sur-Yvette, Île-de-France, France
[32]The Institute for Medical Information Processing, Biometry, and Epidemiology, LMU Munich, Munchen, Bayern, Germany
[33]Nieuwenhuis Institute for Educational Research, University of Groningen, Groningen, The Netherlands

AS, 0000-0003-0031-685X; SH, 0000-0002-1304-8615; BA, 0000-0001-9364-4988; BD, 0000-0002-5979-2781; AD, 0000-0003-3989-9941; RH, 0000-0002-1588-7527; SH, 0000-0002-1304-8615; JI, 0000-0003-3118-6859; GN, 0000-0001-5273-0150; DvR, 0000-0002-5030-4091; DRS, 0000-0002-4600-6323; WV, 0000-0002-5855-3885

Many-analysts studies explore how well an empirical claim withstands plausible alternative analyses of the same dataset by multiple, independent analysis teams. Conclusions from these studies typically rely on a single outcome metric (e.g. effect size) provided by each analysis team. Although informative about the range of plausible effects in a dataset, a single effect size from each team does not provide a complete, nuanced understanding of how analysis choices are related to the outcome. We used the Delphi consensus technique with input from 37 experts to develop an 18-item subjective evidence evaluation survey (SEES) to evaluate how each analysis team views the methodological appropriateness of the research design and the strength of evidence for the hypothesis. We illustrate the usefulness of the SEES in providing richer evidence assessment with pilot data from a previous many-analysts study.

1. Introduction

Researchers adopt a wide range of approaches when analysing data, and their equally justifiable choices about statistical procedures, data processing and the inclusions of covariates can affect the conclusions they draw [1,2]. In fields ranging from epidemiology to psychology to economics, concerns have been raised about the robustness of published evidence since researchers find different answers to the same research question with the same data. This uncertainty in the statistical outcomes is not addressed within standard statistical inference practices and usually remains hidden from view when only a single analysis is presented (e.g. [3]), resulting in overconfidence and model myopia [4–7].

The robustness of an empirical claim on the basis of a single (new, preregistered) dataset can be assessed through multiverse or vibration of effects analysis [8,9] and many-analysts approaches [6] (but see [10] for a critical reflection on robustness analyses). These approaches are designed to reveal the range of justifiable analytic decisions and their consequences for the reported outcome. In a multiverse or vibration of effects analysis, different analytic paths are systematically explored by the same analyst(s) (e.g. [9,11–15]). In a many-analysts project (also referred to as 'crowdsourced analyses'

[6] or a 'multi-analyst approach' [4,7]), different independent analysis teams analyse the same dataset (e.g. [16–23]). In both cases, the end result is an evaluation of the consistency of the observed outcomes across all analyses.

Many-analysts projects appear particularly well suited to mitigate arbitrariness of individual analytic choices, while still allowing for expertise-based analytic decisions concerning data preprocessing, variable exclusion, and model specification. By drawing from a pool of plausible analyses, a many-analysts approach thus enables one to quantify variability across teams based on theory-driven analysis choices and plausible statistical models rather than emphasizing just one analyst's approach. Specifically, if a range of different experts arrive at the same conclusion, we can be fairly confident that the effect is robust. If they reveal a wide variety of outcomes, we need to evaluate why those choices matter.

Many-analysts projects are a recent innovation, but they have already been adopted in many different fields, including neuroscience [16,17,24–26], economics [27,28], epidemiology [29], ecology [30,31], political science [18] and psychology [19–23,32–35]. Many of these projects concluded that different but justifiable analytic decisions led to diverging outcomes, sometimes with statistically significant effects in opposite directions (e.g. [18,28,34], but see [23]).

1.1. Beyond effect sizes: acknowledging insights and concerns of analysis teams

The many-analysts approach can reveal the extent to which the reported outcome varies with different, expert-driven analytical decisions. The approach typically focuses exclusively on a single outcome of interest from each team (such as an odds ratio (e.g. [21]) or a standardized beta coefficient; e.g. [23], but see [17]). These effect size estimates are (visually) summarized to provide an overall impression of the results (but see [36,37] for recently proposed alternative statistical approaches).

This exclusive focus on effect size estimates from each team carries several implicit assumptions: (a) the statistical analyses of each team are sufficiently similar so that they can be summarized using a common effect size metric, (b) further insights from the analysis teams are not relevant when measuring the consistency of the reported results, and (c) analysis teams, by participating in the project, fully endorse the quality of the data they are given and the appropriateness of the research design (cf. 10]).

Commentaries on the recently published Many-Analysts Religion Project [23], studying the relationship between self-reported well-being and religiosity, challenge all three assumptions (see also [18,20,38–51]). First some analysis teams applied more complex approaches that did not naturally yield the specified outcome measure (i.e. standardized regression coefficients). These analyses included structural equation modelling machine learning and even multiverse analyses.[1] Second many teams presented more nuanced interpretations of the primary effect based on sub-group analyses or multivariate approaches which helped determine the conditions under which the hypothesized relation occurred. Third some teams raised concerns about measurement invariance in the data themselves. Others criticized the formulation of the research question an issue that surfaced in the previous many-analysts projects. In sum relying on a single reported effect from each team leaves no room for a more nuanced and detailed interpretation of the results and the underlying data.

1.2. Assessment of subjective evidence

Although measuring the distribution of plausible effect sizes can provide important insights about the robustness of an empirical result [36,37], we argue that it is incomplete (see also [51,52]). To reap the full benefits of involving multiple analysts, we should also examine the broader context in which analysts made their choices: their prior beliefs about the effect, their assessment of the adequacy of the design, or the stability of the effect; thus a subjective measure of evidence. Here, we define subjective evidence as the extent to which one believes in the presence of the effect or relationship given the data and study design.

The idea of collecting a subjective assessment of research evidence in a systematic, reflective and standardized manner is uncommon in the quantitative social and behavioural sciences. Perhaps one could view the discussion section of an empirical article as a narrative subjective evaluation of the obtained evidence, as this is typically where authors discuss the limitations and implications of the quantitative results. It would be challenging, however, to include a narrative summary for every team in a many-analysts project. By contrast, the measurement of subjective evidence has been

[1]Alternative approaches for synthesizing outcomes in many-analysts projects (e.g. considering only the sign of the effect size; focusing on evidential measures such as *p*-values or Bayes factors) do not seem satisfactory, especially when quantifying the size of the effect is essential (e.g. [51]).

included in previous many-analysts projects, although typically only as single items and not with the intention of capturing different aspects of scientific evidence. For instance, analysts may be asked about the plausibility of the research claim (e.g. [21]) or whether the assumed effect or relation was confirmed by the data (e.g. [23]). Other aspects, such as the stability of the effect or the pertinence of the effect size, remain unexplored. Moreover, while some previous studies included measures regarding the appropriateness of analytic approaches, this was not conducted in a self-reflective manner; typically, the analytic teams assessed each other's analytic strategy. Finally, analysis teams' concerns about the research design and data have been raised through personal correspondence [21] and/or commentaries [18,23], but are not systematically addressed in the manuscript itself. Here, we propose a short, simple and systematic assessment of each team's subjective evaluation of the evidence, design and data.

Other fields commonly use the subjective evaluation of the evidence as a scientific assessment, often to systematically integrate evidence from different studies and sources. In the evaluation of randomized controlled trials and systematic literature reviews, for instance, subjective assessment of evidence is particularly relevant, as objective quantification is difficult. For such reviews, existing guidelines help streamline how authors should evaluate the strength of the evidence, the quality of the study design, and the relevance of the results to answering the research question [53–55]. In addition, subjective assessment of evidence plays a central role when evaluating qualitative research, for instance, to inform the development of guidelines and the formulation of policy [56,57].

Systematic guidelines help define the criteria for subjective evaluations, such as the relevance and adequacy of data, coherence of results, or methodological limitations of the study design. Such a standardized approach would be especially useful for many-analysts projects. Many-analysts projects share similarities with systematic literature reviews, as both require integrating multiple sources to address a single research question. We argue that analysis teams will be able to assess the evidence derived from their analysis more comprehensively if they use criteria similar to those used to assess evidence from randomized controlled trials, systematic reviews, and qualitative research.

1.3. Current study

The aim of the current project is twofold. First, we aimed to advance many-analysts studies by developing a subjective evidence evaluation survey (SEES). This survey includes aspects of evidence covered in the previous literature on subjective scientific assessments. Second, we aimed to develop a methodological and analytic strategy to effectively synthesize responses to the SEES.

The methods proposed here are particularly relevant for project leaders of many-analysts studies. Project leaders can use our methods to capture the beliefs of the analysis teams about the evidence for the hypothesized effect of interest more comprehensively. Furthermore, the SEES identifies potential methodological concerns of the analysis teams and may therefore safeguard against unwarranted certainty in drawing conclusions in many-analysts studies. Importantly, the SEES is intended to supplement—not supplant—objective measures of evidence such as the summary of outcome metrics.

In the following, we will describe the reactive-Delphi expert consensus procedure used in the collaborative development of the SEES. We then present the SEES and illustrate how to use it with responses from analysts in the Many-Analysts Religion Project. Appendix A provides more comprehensive guidance and detailed instructions for using the SEES. A Qualtrics template for the SEES is available on the OSF at: https://osf.io/axg2y.

2. Development of the subjective evidence evaluation survey

The idea for the SEES arose from the experience some of us (A.S. and S.H.) had in leading a many-analysts project (e.g. [23]), in which we felt we lacked the tools to fully and systematically represent the analysis teams' efforts and insights that were privately communicated to us. To that effect, we considered which aspects of subjective evidence would be important to capture systematically and also agreed that the development of such a tool would only be successful if it was developed in collaboration with other experts. For these reasons, we decided to develop the SEES together with an expert panel in relevant scientific areas following a preregistered 'reactive-Delphi' expert consensus procedure [58] as implemented in [7] and [59]. The Delphi procedure iteratively determined the consensus of experts on the selection, wording, and content of items in multiple rounds. The

development of the SEES included the creation of the initial item list, the consensus building using the Delphi method, and a final discussion round to finalize the survey.

2.1. Creating the initial item list

During the planning phase, authors A.S., S.H. and E.J.W. drafted an initial item list containing 22 items, which was based on checklist and guidance articles on systematic literature reviews [55] and evaluating qualitative evidence [57,60], and items used in previous many-analysts studies [21,23].[2]

2.2. Recruitment of the expert panel

On 25 November 2022, we contacted 93 experts, including project leaders of many-analysts and multiverse studies listed in [7], along with co-authors of the same publication. In addition, we reached out to experts in systematic literature reviews and evaluating qualitative evidence (e.g. co-authors of [53–57]). Furthermore, we invited measurement and general methodology experts, selecting them based on our knowledge of publications on cautionary notes and common pitfalls in scale construction, and on Bayesian methodology. Finally, we included experts recommended by fellow panel members. From the 93 experts, 45 agreed to participate in developing the SEES, seven declined our invitation, 38 did not respond to our request and three invitations bounced. Of these 45 experts, 37 finished all three consensus rounds.

2.3. Expert consensus procedure

We conducted a total of three rounds of rating by the Delphi method. In each round, the experts rated each item on a 9-point Likert-type scale ranging from 1 (Definitely not include this item) to 9 (Definitely include this item). Based on the panel responses, we iteratively refined our survey in each round by deleting, adding, or rewording items until we achieved consensus and support.

We preregistered a criterion that items with a median recommendation rating of 6 or higher and an interquartile range of 2 or smaller (indicating consensus) would be eligible for inclusion in the SEES (cf. [7,59]).[3] This criterion was applied to all items except one. In round 3 of the expert consensus procedure, item 8 from the subjective evidence subscale received a median support rating of 8 but lacked consensus, with an interquartile range of 4. Despite the large interquartile range, we chose to add this item to the survey based on its high level of support. Specifically, the discussion round and feedback from the panel members suggested that the majority felt quite strongly that the item should be included as it addressed an aspect of evidence that was otherwise lacking (i.e. it was not covered by other items yet); 70% of the experts strongly endorsed the item and gave a score of 7–9 on the inclusion scale. All items received approval from panel members during the discussion round.

We also polled the panel on whether to offer a dichotomous response option or a Likert-type scale for each item in the SEES. The panel preferred a Likert-type scale. In the final version of the SEES, the items are answered using a 4-point Likert scale. The ideal number of response options is a topic of debate (see also [61] for a comprehensive review of the literature concerning survey and scale construction). When comparing 5-point Likert scales to 7-point Likert scales and 10-point Likert scales, some research suggests that a higher number of response options may possess better psychometric qualities (e.g. [62]), while others indicate that a lower number of answer options (5-point Likert scales versus 7-point Likert scales) yields comparable data quality (e.g. [63]). Ultimately, we chose to include a smaller number of answer options to keep the response options clear and interpretable. Additionally, we chose to omit a midpoint as a response option. Scales without midpoint and with a smaller number of answer options (i.e. 2 or 4) demonstrate in some research similar or slightly higher reliability compared to scales with a midpoint (e.g. [64]). Apart from considering psychometric properties, we also wanted to discourage participants from choosing the midpoint when they are undecided [65–67] or when an item is not applicable to their analyses. For such cases, we provided an 'not applicable/I do not know' answer option. A detailed description of the different stages of the Delphi method can be found at https://osf.io/jk674/.

[2]The initial item list can be accessed via https://osf.io/jk674/.

[3]Note that a median rating of 6 is lower than the typical median rating of 8 in medical sciences, where Delphi procedures are frequently applied. We adopted a more lenient cutoff as methodological recommendations for the social and behavioural sciences differ somewhat from those aimed at reaching a medical consensus, where even slight hesitation warrants caution.

3. The subjective evidence evaluation survey

The final version of the SEES consists of 18 items divided across two subscales asking (a) how their beliefs in the hypothesized effect of the study changed after their analyses ('subjective evidence subscale') and (b) whether they thought the methodology of the study was appropriate ('methodological appropriateness subscale').

The full survey is intended to be administered after analysis teams have conducted their analyses (but before they have seen other teams' findings) and submitted their conclusions to the project leaders. In case analysis teams consist of multiple researchers, the survey should be filled out once per team.

3.1. Prior beliefs on plausibility

We recommend asking analysis teams how they evaluate the plausibility of the hypothesis of interest (i.e. item 1 of the subjective evidence subscale) *before having seen the data*. This not only provides valuable information on how the hypothesis is perceived, but also allows the project leaders to investigate confirmation bias (i.e. are prior beliefs related to reported outcomes?) and belief updating (i.e. are posterior beliefs related to reported outcomes and/or is the shift in beliefs related to the reported outcomes?). This item could be embedded in a questionnaire that captures the background and demographic information of the analysis teams (e.g. their expertise, academic position, familiarity with the topic).

1. Before having seen the data, do you find the hypothesized effect or relation plausible?

This item is answered on a 4-point Likert scale with response options 'yes, definitely', 'yes, mostly', 'no, mostly not', and 'no, definitely not'. Project leaders can choose whether or not to include a 'not applicable/I do not know' option. This option is probably not necessary when assessing prior beliefs, but it could be added for consistency with the wording of items in the subjective evidence subscale.

3.2. Subjective evidence subscale

The subjective evidence subscale consists of eight questions, each accompanied by a short example (in italics) to illustrate the intended meaning. All items are answered on a 4-point Likert scale with response options 'yes, definitely', 'yes, mostly', 'no, mostly not', 'no, definitely not', and a 'not applicable/I do not know' option. Counter-indicative items (i.e. items 4, 5, 6 and 7 that indicate lower belief in the hypothesis) are to be reverse-coded. Analysis teams can provide additional feedback for each item in an open text box.

3.2.1. Questions

SE_1 [Plausibility] Taking into account the results of your analyses, do you find the hypothesized effect or relation plausible? *For instance, obtaining substantial evidence that forcing a smiling facial position increases funniness ratings of cartoons shifts your beliefs on the facial feedback hypothesis from sceptical to favourable.*

SE_2 [Robustness] If applicable, is the hypothesized effect or relation consistent across all conducted analyses? *For instance, results from robustness checks or sensitivity analyses are consistent with the hypothesized effect found in the primary analysis.*

SE_3 [Evidence for effect] Does your analysis based on the observed data provide substantial evidence for the hypothesized effect or relation? *For instance, in a study on the recognition speed of words versus non-words, the confidence/credible interval of the effect size does not include zero.*

SE_4 [No evidence against effect*] Does your analysis based on the observed data provide substantial evidence *against* the hypothesized effect or relation? *For instance, evidence points in the opposite direction than hypothesized, or the evidence favours the null hypothesis.*

SE_5 [Subgroup homogeneity*] If applicable, does the hypothesized effect or relation vary between subgroups or data exclusion criteria? *For instance, a treatment benefited patients with moderate or severe depression but not patients with mild depression.*

SE_6 [Subconstruct homogeneity*] If applicable, does the hypothesized effect or relation vary for the different facets of the construct? *For instance, in a study on religiosity and well-being, religiosity was related to psychological and social well-being but not to physical well-being, that is, the relation is not stable across all measured facets of the variable well-being.*

SE_7 [No alternative explanations*] Do your analyses suggest plausible alternative explanations for the hypothesized effect or relation? *For instance, including socioeconomic status as a covariate eliminates the hypothesized relation between place of residence (rural versus urban) and happiness.*

SE_8 [Substantial effect size] Do you believe the size of the effect is substantial enough to be translated into real-life implications? *For instance, an effect of 2 points on a 7-point happiness scale might be perceived as having real-life consequences, whereas an effect of 0.1 points might not.*

3.3. Methodological appropriateness subscale

The methodological appropriateness subscale consists of 10 items, each accompanied by a short example (in italics) to illustrate the intended meaning. All items are answered on a 4-point Likert scale with response options 'major concerns', 'moderate concerns', 'minor concerns', 'no concerns', and a 'not applicable/I do not know' option. Analysis teams can provide additional feedback for each item in an open text box.

3.4. Questions

MA_1 [Sampling plan] Do you have concerns about the appropriateness of the sampling plan for the objectives of the research? *For instance, a study on global religiosity was conducted only in countries that are predominantly Christian which is a threat to external validity.*

MA_2 [Statistical power] Do you have concerns that the number of observations may not be sufficient to assess the hypothesized effect or relation? *For instance, there were not enough trials within participants or participants in conditions to reach sufficient statistical power.*

MA_3 [Missing values] Do you have concerns about missing values on the relevant variables? *For instance, there are too many missing values to draw a statistically valid conclusion, or the pattern of missing values appears non-random.*

MA_4 [Biased sample] Do particular sample characteristics (e.g. age, gender, socioeconomic status) raise concerns for the hypothesized effect or relation? *For instance, in a study on cognitive decline, the average age of the sample of older adults was relatively low (e.g. 60 years), which is a threat to generalizability across populations.*

MA_5 [Study setting] Do particular characteristics related to the setting of the study raise concerns for the hypothesized effect or relation? *For instance, a study on live social interactions was researched online, which is a threat to generalizability across contexts.*

MA_6 [Reliability] Do you have concerns about the reliability of the primary measures (i.e. measures producing similar results under consistent conditions)? *For instance, the measures were internally inconsistent, that is, results across items measuring a given construct were not consistent as indicated by Cronbach's alpha.*

MA_7 [Validity] Do you have concerns about the validity of the measures (i.e. whether the measures capture the constructs of interest)? *For instance, a person's level of social skills was measured by the number of friends they have, which is a threat to construct validity.*

MA_8 [Research design] Do you have concerns about the appropriateness of the research design for addressing the aims of the research? *For instance, a correlational study on obesity and depression was conducted to determine whether obesity causes depression.*

MA_9 [Missing variables] Do you have concerns that some necessary variables were missing to assess the hypothesized effect or relation? *For instance, a pre-intervention baseline measure, a control group, or important covariates were missing.*

MA_10 [Analysis] Do you have concerns about the appropriateness of your analysis for answering the research question? *For instance, some statistical assumptions were violated and could not be sufficiently addressed in the analysis.*

3.5. Computational model

The computational model we developed to synthesize responses to the SEES is based on cultural consensus theory [68–71]. Cultural consensus theory models are used in the analysis of response data where there is no 'ground truth' but the goal is to determine a collective opinion on a specific topic. Applied to the SEES for many-analysts studies, the cultural consensus theory model estimates the analysis teams' collective opinion for each of the scale items, henceforth referred to as *consensus*, as well as overall, on the evidence in the data related to the research question.

We believe the Bayesian cultural consensus theory model is preferable to standard descriptive statistics such as sum scores or means given the structure of the data and desiderata for the results. First, the model is suited for ordinal data, which is not straightforward in many measurement/simple hierarchical models. Second, the model is relatively simple and closely related to well-known IRT models [69] such as the Graded Response Model [72] and the Item Factor Model [73]. Third, the model is hierarchical on the item level, assuming a shared 'latent' variable per subscale and taking into account similarities between analysis teams. Fourth, in addition to the hierarchical estimation of the consensus answer, the model also estimates a between-analysts parameter (i.e. scepticism, defined as an analyst's tendency to select lower (versus higher) values on the response scale) and a between-item parameter (i.e. 'difficulty'; whether the items elicit polarizing responses from the analysts) which may carry relevant information for understanding the results. Fifth, the Bayesian approach (a) enables estimation even with a small number of observations (which is likely in many-analysts projects), while the central limit theorem may fail to provide reliable estimates and (b) provides a quantification of uncertainty by means of the credible intervals [74–76]. Sixth, the model allows for extension to investigate different cultures of raters, which may be of interest in some many-analysts projects (e.g. compare evidence evaluation by theoretical experts to methodological experts).

Specifically, the applied computational model is an adapted version of the latent truth rater model proposed in [69] and extended in [76]. The model is implemented in the Stan programming language using the No-U-Turn sampler [77,78]. Appendix B contains a formal description of the model. In addition, appendix B contains a validation of the two-component structure of the SEES using Bayesian confirmatory factor analysis for ordinal data using the `blavaan` package in R [79].

3.6. An example application of the subjective evidence evaluation survey

To showcase the intended use of the SEES in a many-analysts project, we asked the analysis teams ($N = 120$) of the Many-Analysts Religion Project to retrospectively fill out the preliminary version of the SEES (i.e. the round 3 version of the expert panel procedure; see https://osf.io/4ypzv) based on their analysis for the project's first research question: 'Do religious people self-report higher well-being?', approximately one year after the project had been completed. For this research question, all but three teams reported positive effect size estimates (standardized beta coefficients) with confidence/credible intervals excluding zero, suggesting a positive relation between religiosity and self-reported well-being in the dataset.

The SEES survey was completed by 42 analysis teams (35% of all analysis teams) and therefore these data cannot be taken to reflect the overall consensus from the Many-Analysts Religion Project. The sample of responders does not appear to be biased with regard to self-reported expertise and reported effect sizes. That is, the overall median and its median absolute deviation of the reported effect sizes in the sample of non-responders in the Many-Analysts Religion Project (0.114 [0.035]) are comparable to those of our subsample (0.129 [0.044]). Additionally, responders and non-responders are similar regarding the means and standard deviations of their self-reported methodological knowledge ($M = 4.07$, s.d. = 0.64 for responders versus $M = 4.01$, s.d. = 0.71 for non-responders) and substantive knowledge ($M = 2.76$, s.d. = 1.41 for responders versus $M = 2.63$, s.d. = 1.22 for non-responders). Nevertheless, these data merely serve as an illustration for how to use and analyse the SEES.

For each team, we assessed (1) the collective opinions for each survey item as well as the overall collective opinion for both subscales, (2) the change from prior to final beliefs about the plausibility of the effect, and (3) the correlations between the reported effect sizes and the prior beliefs, final beliefs, and the estimates of individual scepticism. Note that the Many-Analysts Religion Project analysis teams filled out a preliminary version of the SEES (i.e. the version after the third and final round of the consensus procedure) in which the subjective evidence subscale was phrased as statements rather than questions. In the final discussion round these items were reworded as questions rather than statements for consistency with the methodological appropriateness subscale. The response scale labels were changed from 'strongly agree', 'somewhat agree', 'somewhat disagree', and 'strongly disagree' to 'yes, definitely', 'yes, mostly', 'no, mostly not', 'no, definitely not'. We avoided modifying other aspects of the items, including their content. As an illustration, the first item in the subjective evidence subscale (i.e. SE_1 [Plausibility]) was worded as follows in the preliminary version of the SEES provided to the 42 analysts: 'Taking into account the results of my analyses: I find the hypothesized effect or relation plausible. *For instance, having obtained substantial evidence that forcing a smiling facial position increases funniness ratings of cartoons shifts your beliefs on the facial feedback hypothesis from skeptical to favourable*'.

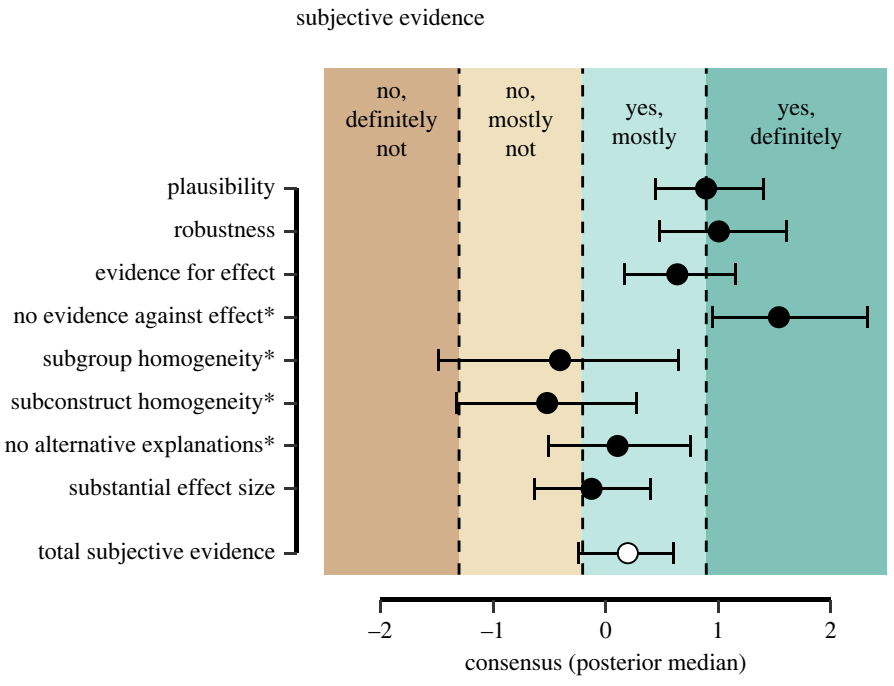

Figure 1. Estimated consensus for the subjective evidence subscale. The black points show the posterior medians (plus 95% credible interval) of the consensus, including the category thresholds. Items followed by an asterisk reflect items that have been reverse-coded and their labels have been changed for interpretability. The white marker at the bottom reflects the overall median assessment (plus 95% CI) of the subjective evidence subscale.

3.6.1. Subjective evidence

Figure 1 shows the model-based consensus for each item of the subjective evidence subscale, including the average response category thresholds and their labels (see also figure 11 for a depiction of the observed item ratings). The consensus represents the true location of the items on an assumed underlying unidimensional scale ranging from minus to plus infinity. When interpreting the consensus, it is crucial to note that they can only be interpreted in relation to the response category thresholds, which are represented by different colours in the figure. The posterior median and 95% credible interval of the overall consensus for the *subjective evidence* subscale is 0.20 [−0.24, 0.60] (visualized by the white marker in the figure) and thus largely falls into the response category 'Yes, mostly', and the standard deviation across items is 0.71. This indicates that the general consensus is that the analysis teams mostly believe that their analysis provides evidence for the hypothesis that religious people self-report higher well-being, with some variation across items. For instance, for item 4 ('no evidence against effect') the majority of analysis teams indicated that there is no evidence against the effect of interest (i.e. they indicated that '[their] analysis based on the observed data did definitely not provide substantial evidence *against* the hypothesized effect or relation'). This result is in line with the fact that all effect sizes were positive, which led the authors to conclude that in this dataset indeed there is a relation between religiosity and self-reported well-being.

For items 5 ('subgroup homogeneity') and 6 ('subconstruct homogeneity'), the analysis teams seem neutral regarding whether effects differed across subgroup analyses and whether effects differed across different subconstructs. The large credible intervals for these items reflect the uncertainty in the estimated consensus, as more than half of the analysis teams considered the items not applicable (suggesting that they did not conduct subgroup analyses or subconstruct analysis). Despite the great uncertainty, however, these items contain noteworthy information. The teams that conducted these analyses seem relatively sceptical about the effect, especially compared to the other aspects of evidence. The cautious scepticism regarding subconstruct homogeneity aligns with the findings of the team leaders from the Many-Analysts Religion Project, as well as some of the published commentaries. Specifically, the project leaders indicated that the relationship between religiosity and well-being showed the strongest relationship when focusing on psychological well-being, followed by social well-being, and least with physical well-being [50]. This insight was gained through exploring the teams' reported operationalizations of the dependent variable. However, scepticism regarding

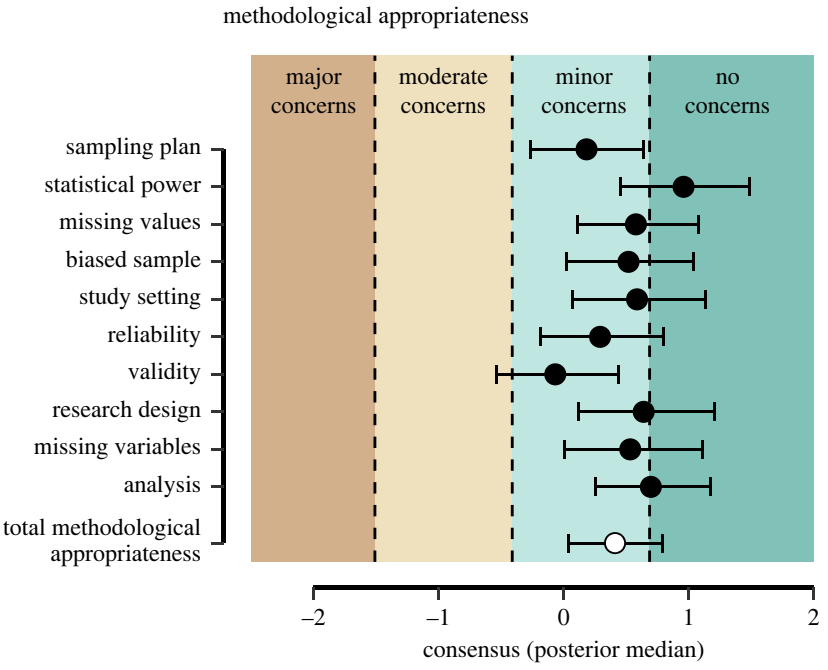

Figure 2. Estimated consensus for the methodological appropriateness subscale. The black points show the posterior medians (plus 95% credible interval) of the consensus, including the category thresholds. The white marker at the bottom reflects the overall median assessment (plus 95% CI) of the methodological appropriateness subscale.

subgroup homogeneity had not been addressed by the project leaders. The implementation of SEES facilitates the identification of such nuanced insights that otherwise remain hidden and prompts team leaders to delve deeper into analytic results underlying sceptical responses.

3.6.2. Methodological appropriateness

Figure 2 shows the model-based consensus for each item of the 10 items of the methodological appropriateness subscale (see also figure 12 for a depiction of the observed item ratings). Compared to the subjective evidence subscale, the latent consensus for the methodological appropriateness subscale appears more similar. The posterior median of the overall consensus for the *methodological appropriateness* subscale is 0.41 with a 95% credible interval ranging from 0.04 to 0.79 and the standard deviation across items is 0.33. This indicates that the general consensus of the analysis teams is that there are minor to no methodological concerns regarding the analysis of the hypothesis that religious people self-report higher well-being. The posterior medians for almost all items reflect the analysis teams' assessment of 'minor concerns'; with the exception of item 2 (regarding the sufficiency of the number of observations) for which the posterior median reflects 'no concerns' (and perhaps item 10 on the analysis). For item 7 (validity), the analysis teams were most sceptical, with the credible interval of the consensus reaching an assessment of 'moderate concerns'. This may reflect certain concerns raised in the published commentaries of the Many-Analysts Religion Project regarding measurement invariance [47,48], specifically, the observation that the religiosity construct does not maintain the same factor structure across all included countries.

3.6.3. Subjective beliefs and effect size estimates

Figure 3 displays the average prior and final beliefs about the plausibility of the hypothesis of interest.[4] Researchers' prior beliefs about religiosity being positively related to self-reported well-being were already high ($M = 3.00$ on the 4-point Likert scale), but were raised further after having conducted the analysis ($M = 3.48$ on the 4-point Likert scale). Specifically, before seeing the data, 73.81% of the teams

[4]Note that prior beliefs about the plausibility of the effect were reported on a 7-point scale instead of a 4-point scale in the Many-Analysts Religion Project. To make these prior beliefs compatible with the posterior plausibility assessment as included in the SEES (item 6), we recoded the 7-point scale into a 4-point scale: 1 became 1, 3 became 2, 5 became 3 and 7 became 4. Responses in the in-between categories were randomly assigned; 2 was randomly assigned to become 1 or 2, 4 was randomly assigned to become 2 or 3, and 6 was randomly assigned to become 3 or 4.

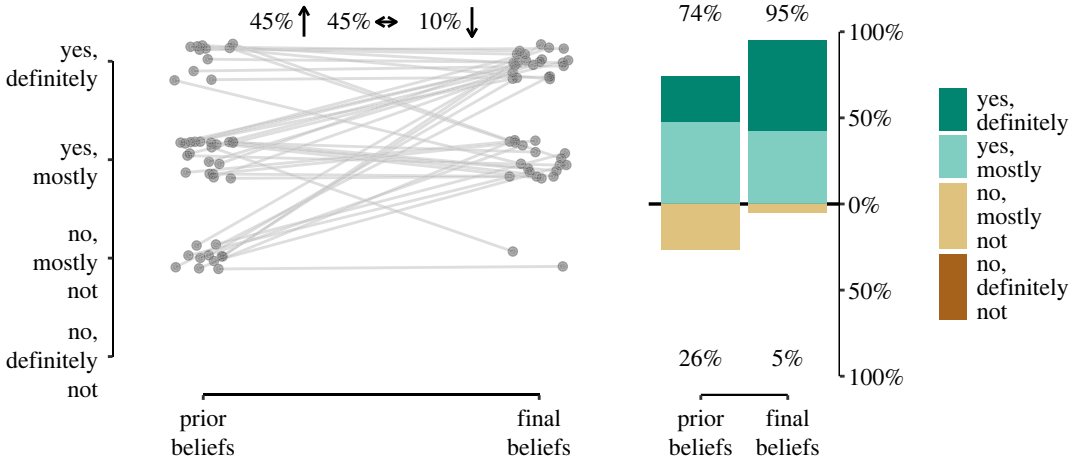

Figure 3. Prior and final beliefs about the plausibility of the hypothesis. The left side of the figure shows the change in beliefs for each analysis team. Forty-five per cent of the teams considered the hypothesis more likely after having analysed the data than prior to seeing the data, 10% considered the hypothesis less likely having analysed the data, and 45% did not change their beliefs. Plausibility was measured on a 4-point Likert scale ranging from 'strongly disagree' to 'strongly agree'. Points are jittered to enhance visibility. The right side of the figure shows the distribution of the Likert response options before and after having conducted the analyses. The number at the top of the data bar indicates the percentage of teams that agreed that the hypothesis was plausible (in green) and the number at the bottom of the data bar (in brown/orange) indicates the percentage of teams that disagreed that the hypothesis was plausible.

considered it likely that religiosity is related to higher self-reported well-being. This percentage increased to 95.24% after having seen the data.

Following [21,23], we explored whether expectations and confirmation bias influenced the outcomes of the analysis teams and whether analysis teams updated their beliefs after conducting their analysis. To this aim, we assessed whether the reported effect sizes were positively related to the subjective assessments of the plausibility of the research question before and after analysing the data. In addition, we evaluated whether the effect sizes were related to the estimates of individual scepticism (i.e. their general tendency to select lower answer options on the scale; corresponding to the β parameters in the formal model description) on the subjective evidence subscale. Here, we would expect a *negative* correlation between effect size estimates and individual scepticism, reflecting that analysis teams who found lower effect sizes were subjectively more sceptical (less optimistic) about evidence for the research question. These hypotheses were tested against the null hypothesis that there is no relation between reported effect sizes and subjective beliefs or scepticism. As subjective beliefs were measured on a 4-point Likert scale, we used a rank-based Spearman correlation for the first two correlations and a Pearson correlation for the relation between effect size and estimated scepticism.

The correlations are visualized in figure 4. We obtained moderate evidence *against* a positive relation between prior beliefs about the plausibility of the hypothesis and the reported effect sizes: $BF_{+0} = 0.13$; $BF_{0+} = 7.94$, $\rho_s = -0.11$, 95% credible interval [−0.38, 0.21]. In addition, we found strong evidence against a positive relation between posterior beliefs about the plausibility of the research question and the reported effect sizes: $BF_{+0} = 0.08$; $BF_{0+} = 12.11$, $\rho_s = -0.27$, 95% credible interval [−0.53, 0.06].[5] Finally, we found anecdotal evidence against a negative relation between estimated scepticism on the SEES and reported effect sizes: $BF_{+0} = 0.35$; $BF_{0+} = 2.85$, $\rho = 0.00$, 95% credible interval [−0.30, 0.28].

As mentioned in [23], these results provide no indication that expectations and confirmation bias influenced the teams' results (i.e. prior beliefs are not related to reported effect sizes), nor do they

[5]In the sample of non-responders from the Many-Analysts Religion Project, we reach the same conclusions regarding the absence of evidence for confirmation bias. That is, we found strong evidence *against* a positive relation between prior beliefs about the plausibility of the hypothesis and the reported effect sizes $BF_{+0} = 0.08$; $BF_{0+} = 12.23$, $\rho_s = -0.12$, 95% credible interval [−0.34, 0.12]. Regarding the updating of beliefs we found inconclusive evidence in the sample of non-responders. The relation between posterior beliefs about the plausibility of the research question and the reported effect sizes was $BF_{+0} = 1.09$; $BF_{0+} = 0.91$, $\rho_s = 0.19$, 95% credible interval [−0.03, 0.38].

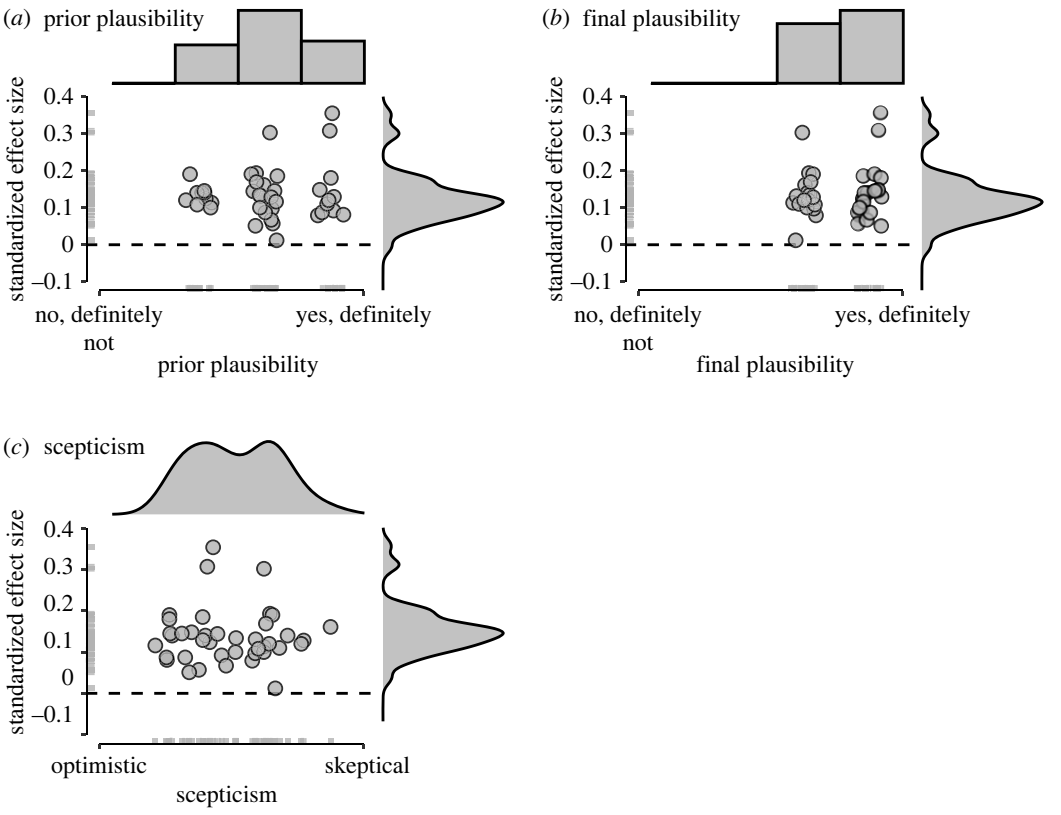

Figure 4. Reported effect sizes (beta coefficients) and subjective beliefs about the likelihood of the hypothesis. Panel (*a*) shows the relation between effect size and prior beliefs for the research question. Panel (*b*) shows the relation between effect size and final beliefs for the research question and panel (*c*) shows the relation between effect size and the analysis teams' level of scepticism regarding the evidence. In (*a*,*b*), points are jittered on the *x*-axis to enhance visibility. The dashed line represents an effect size of 0. Histograms/density plots at the top represent the distribution of subjective beliefs and the density plots on the right represent the distribution of reported effect sizes.

provide evidence for belief updating after having conducted the analyses (i.e. posterior beliefs are not related to the reported effect sizes). Note, however, that the updating of beliefs may not have happened because prior beliefs about research question 1 were already in line with the outcomes; most teams expected and reported evidence for a positive relation between religiosity and well-being, with little variation between teams in both beliefs and reported effect sizes. This lack of variability across teams may also underlie the absence of a correlation between individual differences in objectively reported effect sizes and estimated scepticism. In cases where the analysis teams report diverging results (i.e. conclusions that are qualitatively different) one may expect to find stronger belief updating and larger variability in individual scepticism.[6]

Moreover, the long time period between conducting the analyses and completing the SEES prevents strong interpretations of these results. Instead, as mentioned at the beginning of this section, the data presented here should be regarded merely as a demonstration of the intended use of the SEES.

4. Discussion

The present work introduces the SEES as a tool to systematically explore and quantify subjective measures of evidence in many-analysts projects. The development of the SEES was informed by work on systematic reviews and qualitative research and was collaboratively developed by 37 experts in related fields in a reactive-Delphi procedure, reflecting a consensus among these experts. The 18-item survey covers various aspects of evidence, such as coherence, robustness, and relevance as well as

[6]For the second research question discussed in the Many-Analysts Religion Project, analysis teams reported more qualitatively different results compared to the first research question. And indeed, in this case, we found strong evidence for belief updating, that is, posterior beliefs were positively correlated with reported effect sizes [23].

diverse methodological concerns regarding the underlying design and data that may affect the interpretation of the obtained statistical results.

The first aim of the current project was to develop a measurement tool to capture analysts' beliefs about the evidence obtained in a many-analysts project. Combined with the objective outcomes of the many-analysts approach such as effect size estimates or proportion of statistically significant results, the SEES contributes to a comprehensive summary of the obtained evidence for the hypothesis of interest in a many-analysts project. By capturing analysts' beliefs about the evidence of the hypothesis of interest, the SEES presents a solution to a challenging task: bringing insights and concerns of the analysis teams to the surface in a systematic and scalable manner.

Rather than requiring each team to write a narrative evaluation, project leaders can have them complete the SEES to extract a collective assessment of insights and concerns from all participating teams. Here, we suggested to have the SEES completed once per analysis team. However, if the (additional) goal is to identify potential within-team variability, project leaders may consider eliciting one answer per analyst. This approach may require an extension to the proposed cultural consensus theory model to account for dependencies of analysts within teams.

The SEES introduces several advantages over previously employed methods for subjectively assessing evidence in many-analysts projects. First, in past projects, only a limited number of items concerning subjective evidence (e.g. plausibility of the effect) were administered to analysis teams, primarily to evaluate confirmation bias and belief updating [21]. By contrast, the SEES encompasses a comprehensive list of aspects of subjective evidence and methodological appropriateness deemed important by experts. Second, prior many-analysts studies did not gather information on how analysts themselves rated the appropriateness of their analysis; instead, if analyses were evaluated, the quality of analysis was typically rated by other participating analysts. The SEES allows analysts to reflect on the appropriateness of their own analysis, which can either replace or supplement time-consuming peer assessments. Third, by including the methodological appropriateness subscale, the SEES also offers teams a platform to indicate the quality of the provided dataset and research design. The insights gained from the SEES can encourage team leaders to perform additional analyses (e.g. subgroup or subconstruct analyses) that add important nuances to the main results, as seen in our example application. At minimum, it compels the project leads to reflect on the methodological concerns indicated by the analysts.

Importantly, we do not advocate replacing objective measures of evidence with subjective measures. The subjective measures of evidence complement the objective measures by putting the findings in perspective and/or highlighting inconsistencies in the results or flaws in the research design. The subjective evaluation captured by the SEES provides concrete input for the general discussion of a many-analysts manuscript. In addition, answers to the SEES might reveal potential sources of variability in the obtained results; for instance, teams that investigated different subgroups might reach different conclusions and obtain a different outcome metrics from teams that only targeted one large group.

The second aim of the current project was to outline an analytic strategy for interpreting SEES outcomes, quantifying belief updating in analysis teams, and connecting outcomes of the SEES with objective outcome measures. Concretely, this strategy allows project leaders to investigate whether prior expectations or confirmation bias influenced the results (cf. [21]). We recommend using the outlined Bayesian cultural consensus theory model to analyse the SEES data, but also acknowledge that our analysis strategy is not necessary when employing the SEES. Instead, project leaders may opt to calculate sum scores per subscale and/or overall sum scores for the entire survey (see the appendix for a visualization of the results from the example application based on descriptive statistics).

We contribute to the current literature about guidelines on many-analysts studies [7] by offering concrete advice on how to analyse and interpret (part of) the data obtained in many-analysts projects. This, together with advancements on synthesizing objective outcome metrics across analyses based on the same data (e.g. [36,37]), can move the field beyond drawing conclusions based on (visual) inspection of the analysts' outcomes.

4.1. Limitations

The applicability of the SEES may vary depending on the nature of the many-analysts project. In a typical scenario, many-analysts projects leave some room for analytic flexibility and subjective principled decisions. The nuances that arise from this subjectivity can then be captured by the SEES. There might be situations, however, where many-analysts projects essentially eliminate opportunity for analysts to

choose which variables to assess. For instance, commentaries published on the Many-Analysts Religion Project (e.g. [42,49]) argued that project leaders should have reduced analytic variability arising from differences in the interpretation of the research question. For instance, instead of asking 'Do religious people report higher well-being?' it would be more beneficial to ask whether '*specific* behaviours and/or beliefs benefit *specific* populations' well-being or health in *specific* contexts' [42, p. 2]. In a many-analysts project which aims to reduce analytic variability stemming from different interpretations of the research question as much as possible, capturing the subjective evaluations of the analysis teams may not be advised.

Analysis teams in the Many-Analysts Religion Project, however, viewed the exploration of subgroup effects or of testing of the effect across different sub-constructs as integral to answering the research question. The results from a more generally formulated research question may therefore be more typical of the heterogeneity in the literature on a particular effect. After all, an important motivation to conduct a many-analysts study is to capture (the consequences of) different principled decisions throughout the analytic process.

In addition, some datasets or research designs may render some items irrelevant. For instance, the item on reliability may be irrelevant if the data only feature single item measures. In [21], for instance, the dependent variable was the number of red cards given to dark skinned soccer players. This variable is a single-item measure in which the reliability (e.g. internal consistency) is irrelevant and the SEES item may confuse the participating teams. Although the analysis teams can always indicate 'not applicable/I do not know', the project leaders may also choose to remove these items from the survey.

When multiple research questions are posed, participating teams may find it cumbersome to answer the SEES for each question. If the research questions are answered based on one dataset, rather than on multiple datasets (e.g. stemming from multiple experiments), project leaders could present the methodological appropriateness subscale just once to the teams.

Finally, we invited experts from previous many-analysts studies and experts in the field of systematic literature reviews and qualitative research. However, the framing of the items and the examples given may speak more to researchers from the social and behavioural sciences than to researchers from other areas. If necessary, project leaders from many-analysts studies could use the SEES flexibly and reword the examples to better fit the specific field of study.

It should be noted that if project leaders use a modified version of the SEES (e.g. removing the reliability item if it is irrelevant in their study), it should not be presented as reflecting a consensus approach because removing items may influence the estimation accuracy of the proposed cultural consensus theory model. The performance of the computational model also depends on the number of participating analysis teams with more precise estimation of consensus as the number of teams increases. In simulation studies, we found good model performance based on visual inspection for a sample of $N = 42$—that is, the sample size in our example application—and recommend using at least that many responses when applying the proposed computational model to the SEES. We also found satisfactory model performance based on visual inspection for 20 analysis teams, although with less favourable recovery for true consensus (i.e. wider posterior distributions) compared to the larger sample. The full results can be accessed in the electronic supplementary material at https://osf.io/4cesj.

4.2. Concluding thoughts

In the current project, we collected pilot data to illustrate the intended use and analysis of the SEES, by asking analysts from the Many-Analysts Religion Project to retroactively complete the survey. An obvious next step would be to implement the SEES in a future many-analysts project. We hope to have inspired project leaders of many-analysts projects to consider adding subjective evidence assessment to their future projects and thereby allowing for a more complete evaluation of the outcomes.

5. Disclosures

5.1. Preregistration

Prior to collecting data, we preregistered the full procedure for the reactive Delphi method to develop the SEES on the Open Science Framework at https://doi.org/10.17605/OSF.IO/E4QNY. Any deviations from the preregistration are mentioned in this paper. Note that we also preregistered a procedure for collecting SEES data from the 2023 cohort of a graduate course. However, since only two teams of

students decided to participate, we could not continue this line of data collection. Instead, we decided to contact the analysts from the Many-Analysts Religion Project again and ask them to retroactively fill out the SEES. This latter approach was not preregistered.

5.2. Ethical approval

The expert consensus procedure was approved by the local ethics board of the University of Amsterdam (registration no. 2022-PML-15535). All participants were treated in accordance with the Declaration of Helsinki. Experts who participated in all consensus rounds and approved the final version of the manuscript were given the opportunity to become co-authors of this publication. The collection of pilot data to illustrate how many-analysts studies can be analysed with the SEES was approved by the local ethics board of the University of Amsterdam (registration number: FMG-4376). Researchers who participated in the pilot study received an €8 voucher as compensation.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. Readers can access the preregistration, the materials for the study, the data and the R code to conduct all analyses (including all figures) in our OSF folder at: https://osf.io/jk674/ [80].

The data are provided in electronic supplementary material [81].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. A.S: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization, writing—original draft, writing—review and editing; S.H.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization, writing—original draft, writing—review and editing; D.B.: formal analysis, methodology, software; B.A.: investigation, writing—review and editing; C.A.: investigation, writing—review and editing; T.A: investigation, writing—review and editing; R.B.-N.: investigation, writing—review and editing, N.A.B., investigation, writing—review and editing, A.M.C., investigation, writing—review and editing; B.D.: investigation, writing—review and editing; A.D.: investigation, writing—review and editing; E.I.F.: investigation, writing—review and editing; R.H.: investigation, writing—review and editing; S.H.: investigation, writing—review and editing; F.H.: investigation, writing—review and editing; J.H.: investigation, writing—review and editing; N.H.-K.: investigation, writing—review and editing; J.I.: investigation, writing—review and editing; M.J.: investigation, writing—review and editing; M.K.: investigation, writing—review and editing; E.L.: investigation, writing—review and editing; J.-F.M.: investigation, writing—review and editing; D.M.: investigation, writing—review and editing; A.M.: investigation, writing—review and editing; G.N.: investigation, writing—review and editing; D.R.: investigation, writing—review and editing; M.S.: investigation, writing—review and editing; H.S.-K.: investigation, writing—review and editing; D.S.: investigation, writing—review and editing; D.J.S.: investigation, writing—review and editing; B.A.S.: investigation, writing—review and editing; A.H.S.: investigation, writing—review and editing; B.S.: investigation, writing—review and editing; D.T.: investigation, writing—review and editing; F.T.: investigation, writing—review and editing; E.L.U.: investigation, writing—review and editing; W.V.: investigation, writing—review and editing; J.W.: investigation, writing—review and editing; E.J.W.: conceptualization, formal analysis, investigation, methodology, project administration, supervision, validation, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. No funding has been received for this article.

Acknowledgements. We thank Frédérique Heppel for her support in creating the initial item list of the SEES and helping with data collection. This work was supported by an Amsterdam Brain and Cognition (ABC) project grant to A.S. (ABC PG 22). T.A. was supported in part by NSF IIS-1901386, NSF CAREER IIS-2142794. N.A.B. is supported by the German Research Foundation (DFG; BU 2400/11-1) and by the DFG priority programme 'META-REP: A Meta-scientific Programme to Analyse and Optimize Replicability in the Behavioural, Social, and Cognitive Sciences'. B.D. was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM104420. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. N.N.N.v.D. was supported by the NWO Vici grant no. 181.029. A.D. was supported by the Knut and Alice Wallenberg Foundation, the Marianne and Marcus Wallenberg Foundation and the Jan Wallander and Tom Hedelius Foundation. D.M. is supported by a Vidi grant (VI.Vidi.191.091) from the Dutch Research Council (NWO). A.J.M. is supported by a Vici grant (016.Vici.185.068) from the Dutch Research Council (NWO). D.v.R. is supported by a Vidi grant (016.Vidi.188.001) from the Dutch Research Council (NWO). D.R.S. was supported by a grant from the UK Economic and Social Research Council. D.T. was supported by Marie Sklodowska-Curie grant agreement no. 101023805 under the European Union's Horizon 2020 research and innovation programme. E.L.U. was supported by an R&D grant from INSEAD. J.W. is supported by a VICI grant (VI.C.221.100) from the Dutch Research Council (NWO).

Appendix A. Subjective evidence evaluation survey: instructions to project leaders

The SEES has been developed to facilitate and extend the interpretation of evidence for a given research hypothesis in many-analysts projects. In a many-analysts project, multiple research teams are invited to independently analyse the same data and address the key research question [6,21]. The core idea of a many-analysts approach is to demonstrate the range of justifiable analytic decisions and their consequences in terms of outcomes and conclusions, thereby unveiling the robustness or fragility of the effect of interest. Results from different analysis teams are typically summarized by means of effect size estimates (e.g. odds ratios or beta weights). The SEES has been developed to complement and extend this method, by allowing analysis teams to subjectively reflect on (a) the evidence that their analysis provides for the hypothesis of interest and (b) the quality of the materials and the data. This allows the analysis teams to communicate a more fine-grained evaluation of the evidence obtained through their analysis, yet in a structured manner.

The SEES consists of two subscales, eliciting analysis teams to answer questions about (a) how analysts' beliefs in the hypothesized effect of the study changed after their analyses (subjective evidence subscale) and (b) whether they thought the methodology of the study was appropriate (methodological appropriateness subscale). The full survey should be administered after analysis teams conducted and submitted their analyses. In addition, in order to assess belief updating, item 1 of the subjective evidence subscale should also be administered *prior* to receiving the to-be-analysed data. In case analysis teams consist of multiple researchers, the survey should be filled out once per team.

A.1. Pre-analysis phase

We recommend asking analysis teams to evaluate the plausibility of the hypothesis of interest *before having seen the data*. This not only provides valuable information on how the hypothesis is perceived, but also allows the project leaders to investigate confirmation bias (i.e. are prior beliefs related to reported outcomes?) and belief updating (i.e. are posterior beliefs related to reported outcomes and/or is the shift in beliefs related to the reported outcomes?). This item could be embedded in a questionnaire on the background of analysis teams (e.g. expertise, academic position, familiarity with the topic).

1. Before having seen the data, do you find the hypothesized effect or relation plausible?

Answer options: 'yes, definitely', 'yes, mostly', 'no, mostly not', and 'no, definitely not'. Project leaders can choose whether or not to include a 'not applicable/I do not know' option. This option is probably not necessary for the pre-analysis survey, but it could be added for consistency with the post-analysis survey.

A.2. Post-analysis phase

A.2.1. Subjective evidence subscale

The subjective evidence subscale consists of eight items. Each item contains the question of interest plus a short example to illustrate the intended meaning (in italics). All items are answered on a 4-point Likert scale with response options 'yes, definitely', 'yes, mostly', 'no, mostly not', 'no, definitely not' and a 'not applicable/I do not know' option. Counter-indicative items (i.e. items indicating lower belief in the hypothesis) are to be reverse-coded (i.e. item 4, 5, 6 and 7). Analysts can provide additional feedback for each item in an open text box. An example of how the items and response options could be presented is shown in figure 5.

A.3. Instructions

'Please answer the following questions about your assessment of the *evidence* based on your analysis for [research question]. We understand that this is subjective; there are no correct or wrong answers. Hypothesis: [the hypothesis of interest]. Please base your answers on your interpretation of the analysis conducted by your team'.

A.4. Questions

1. Taking into account the results of your analyses, do you find the hypothesized effect or relation plausible? *For instance, obtaining substantial evidence that forcing a smiling facial position increases funniness ratings of cartoons shifts your beliefs on the facial feedback hypothesis from sceptical to favourable.*

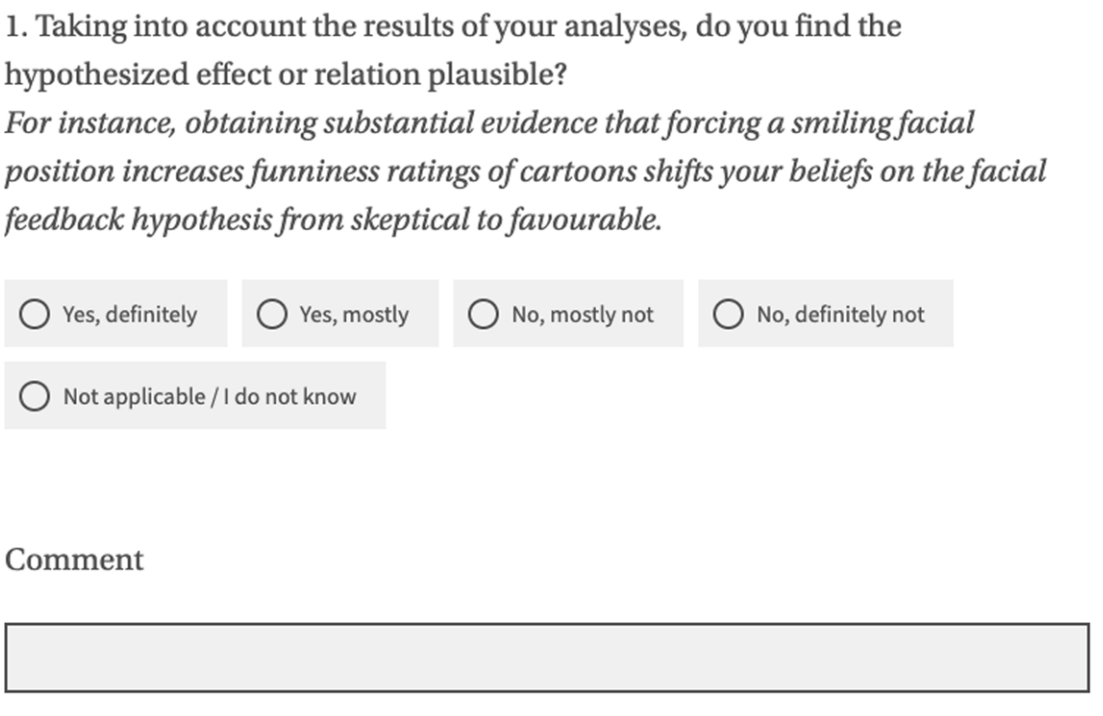

Figure 5. Example of the presentation of the subjective evidence subscale, item 1.

2. If applicable, is the hypothesized effect or relation consistent across all conducted analyses? *For instance, results from robustness checks or sensitivity analyses are consistent with the hypothesized effect found in the primary analysis.*
3. Does your analysis based on the observed data provide substantial evidence for the hypothesized effect or relation? *For instance, in a study on the recognition speed of words versus non-words, the confidence/credible interval of the effect size does not include zero.*
4. Does your analysis based on the observed data provide substantial evidence *against* the hypothesized effect or relation? *For instance, evidence points in the opposite direction than hypothesized, or the evidence favours the null hypothesis.*
5. If applicable, does the hypothesized effect or relation vary between subgroups or data exclusion criteria? *For instance, a treatment benefited patients with moderate or severe depression but not patients with mild depression.*
6. If applicable, does the hypothesized effect or relation vary for the different facets of the construct? *For instance, in a study on religiosity and well-being, religiosity was related to psychological and social well-being but not to physical well-being, that is, the relation is not stable across all measured facets of the variable well-being.*
7. Do your analyses suggest plausible alternative explanations for the hypothesized effect or relation? *For instance, including socioeconomic status as a covariate eliminates the hypothesized relation between place of residence (rural versus urban) and happiness.*
8. Do you believe the size of the effect is substantial enough to be translated into real-life implications? *For instance, an effect of 2 points on a 7-point happiness scale might be perceived as having real-life consequences, whereas an effect of 0.1 points might not.*

A.4.1. Methodological appropriateness subscale

The methodological appropriateness subscale consists of 10 items. Each item contains the question of interest plus a short example to illustrate the intended meaning (in italics). All items are answered on a 4-point Likert scale with response options 'major concerns', 'moderate concerns', 'minor concerns', 'no concerns', and a 'not applicable/I do not know' option. Analysis teams can provide additional feedback for each item in an open text box. An example of how the items and response options could be presented is shown in figure 6.

Instructions. 'Please answer the following questions about your assessment of *methodological concerns* regarding your analysis for [research question]. We understand that this is subjective; there are no correct or wrong answers. Hypothesis: [hypothesis of interest]. Please base your answers on your reflections regarding the provided data and study design'.

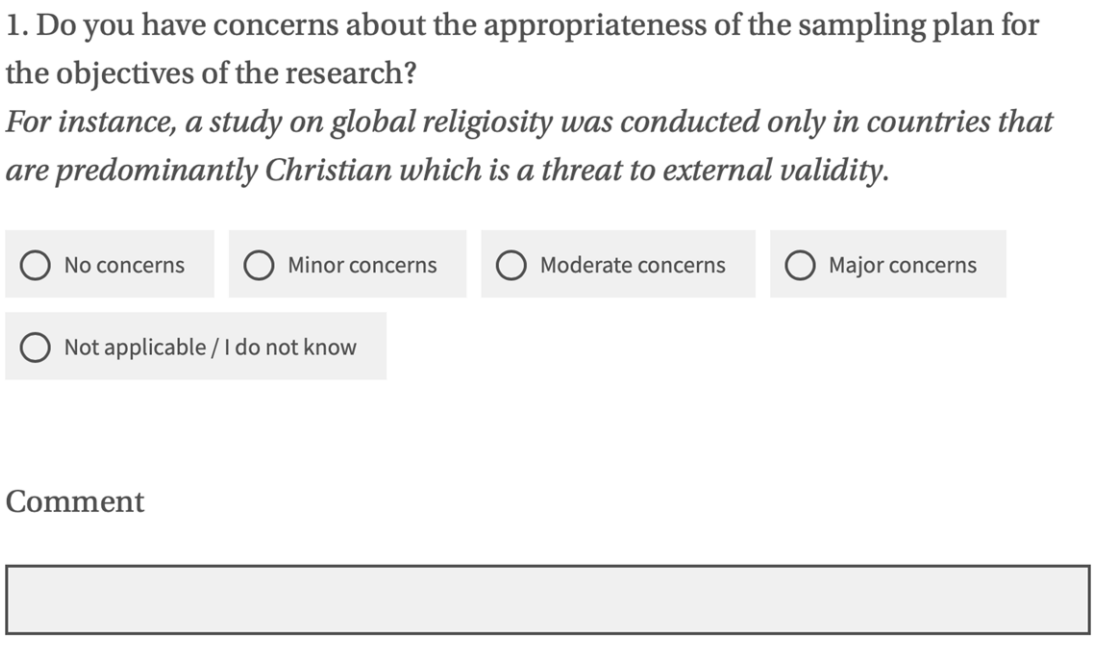

Figure 6. Example of the presentation of the subjective evidence subscale, item 1.

Questions.

1. Do you have concerns about the appropriateness of the sampling plan for the objectives of the research? *For instance, a study on global religiosity was conducted only in countries that are predominantly Christian which is a threat to external validity.*
2. Do you have concerns that the number of observations may not be sufficient to assess the hypothesized effect or relation? *For instance, there were not enough trials within participants or participants in conditions to reach sufficient statistical power.*
3. Do you have concerns about missing values on the relevant variables? *For instance, there are too many missing values to draw a statistically valid conclusion, or the pattern of missing values appears non-random.*
4. Do particular sample characteristics (e.g. age, gender, socio-economic status) raise concerns for the hypothesized effect or relation? *For instance, in a study on cognitive decline, the average age of the sample of older adults was relatively low (e.g. 60 years), which is a threat to generalizability across populations.*
5. Do particular characteristics related to the setting of the study raise concerns for the hypothesized effect or relation? *For instance, a study on live social interactions was researched online, which is a threat to generalizability across contexts.*
6. Do you have concerns about the reliability of the primary measures (i.e. measures producing similar results under consistent conditions)? *For instance, the measures were internally inconsistent, that is, results across items measuring a given construct were not consistent as indicated by Cronbach's alpha.*
7. Do you have concerns about the validity of the measures (i.e. whether the measures capture the constructs of interest)? *For instance, a person's level of social skills was measured by the number of friends they have, which is a threat to construct validity.*
8. Do you have concerns about the appropriateness of the research design for addressing the aims of the research? *For instance, a correlational study on obesity and depression was conducted to determine whether obesity causes depression.*
9. Do you have concerns that some necessary variables were missing to assess the hypothesized effect or relation? *For instance, a pre-intervention baseline measure, a control group, or important covariates were missing.*
10. Do you have concerns about the appropriateness of your analysis for answering the research question? *For instance, some statistical assumptions were violated and could not be sufficiently addressed in the analysis.*

A.5. Background information

A.5.1. Survey development

The SEES was developed in collaboration with 37 experts in relevant scientific areas following a preregistered 'reactive-Delphi' expert consensus procedure [58] as implemented in [7,59]. Selected areas of expertise

included many-analysts and multiverse studies, systematic literature reviews, questionnaire development, and general methodology. Over three rounds, experts were asked to rate each item in the SEES on a 9-point Likert-type recommendation scale ranging from 1 (*Definitely do not include this item*) to 9 (*Definitely include this item*). Based on the panel responses, the survey was iteratively refined in each round by deleting, adding, or rewording items until achieving consensus and support.

We considered items to have reached panel consensus if the interquartile range of expert ratings was 2 or smaller, and we regarded items as having obtained panel support if their median ratings were 6 or higher. Note that we preregistered that only items with a median recommendation rating of 6 or higher and an interquartile range of 2 or smaller would be eligible for inclusion in the SEES. This criterion was applied to all items except one. In round 3 of the expert consensus procedure, item 8 from the subjective evidence subscale received a median support rating of 8 but lacked consensus, with an interquartile range of 4. Despite this, we chose to add this item to the survey. In the discussion round, some panel members explicitly approved the items; none of the panel members objected to them.

A.5.2. Adaptations

We encourage project leaders from many-analysts studies to use the SEES flexibly and reword the examples for the items to the specific field of study if necessary. In addition, some datasets or research designs may render some items irrelevant and may confuse the participating teams. Although the analysis teams can always indicate 'not applicable/I do not know', the project leaders may also choose to remove these items from the survey. Finally, when analysis teams have answered multiple research questions based on one dataset, rather than on multiple datasets (e.g. stemming from multiple experiments), project leaders could present the methodological appropriateness subscale only once to the teams.

A.5.3. Proposed analysis

The SEES data can be analysed in a cultural consensus theory model [68–71,76], which allows one to synthesize responses and capture the collective opinion about the subjective evidence in many-analysts projects. A comprehensive description of the cultural consensus theory model applied to the SEES can be found in appendix B.

Appendix B. Cultural consensus theory model

Here, we describe the adapted version of the latent truth rater model [69,76] which we use to synthesize responses and capture collective opinions on the two subscales of the SEES.

B.1. Model description

Within the latent truth rater model the variable $x_{r,i,s}$ denotes the observed and discrete responses provided by analyst r for item i on subscale s.[7] For convenience, we will drop the subscript s in further descriptions. The variable $x_{r,i}$ takes on the value $x_{r,i} = 0$ when the response to an item is 'not applicable/I do not know'. In all other cases, for each item, it takes on one of the $C = 4$ Likert scores. For instance, in the subjective evidence subscale, $x_{r,i} = 1$ corresponds to the analyst's response of 'no, definitively not', $x_{r,i} = 2$ to 'no, mostly not', $x_{r,i} = 3$ to 'yes, mostly', and $x_{r,i} = 4$ corresponds to 'yes, definitely'.

The model determines three factors that influence the observed responses. The first factor is the applicability of the item which is captured by the probability π_i. Higher values of π_i indicate higher non-applicability resulting in more analysts selecting the 'not applicable/I do not know' option. The remaining responses are influenced by the second and third factors. The second factor, the latent appraisal for the item $y_{r,i}$, is a combination of item properties, such as the latent consensus and the item difficulty (i.e. extent of eliciting polarizing responses). The third factor relates to individual characteristics of the analysts, that is, their individual bias, which determines their decision criteria, denoted as $\delta_{r,c}$.

The latent truth rater model assumes that each item has some latent consensus (originally termed 'item truth' θ_i among analysts—their true collective opinion—on an abstract psychological scale, e.g. the conceived plausibility of the hypothesized effect or relation). Given that an analyst has sufficient knowledge to answer the item, the following process is assumed to take place. Across all items, the analyst draws a mental sample for their decision criteria $\delta_{r,c}$. Then, for each item, they draw a mental

[7]Note that we use the term 'analyst' to refer to the independent analysis teams, which can consist of one or more analysts in practice.

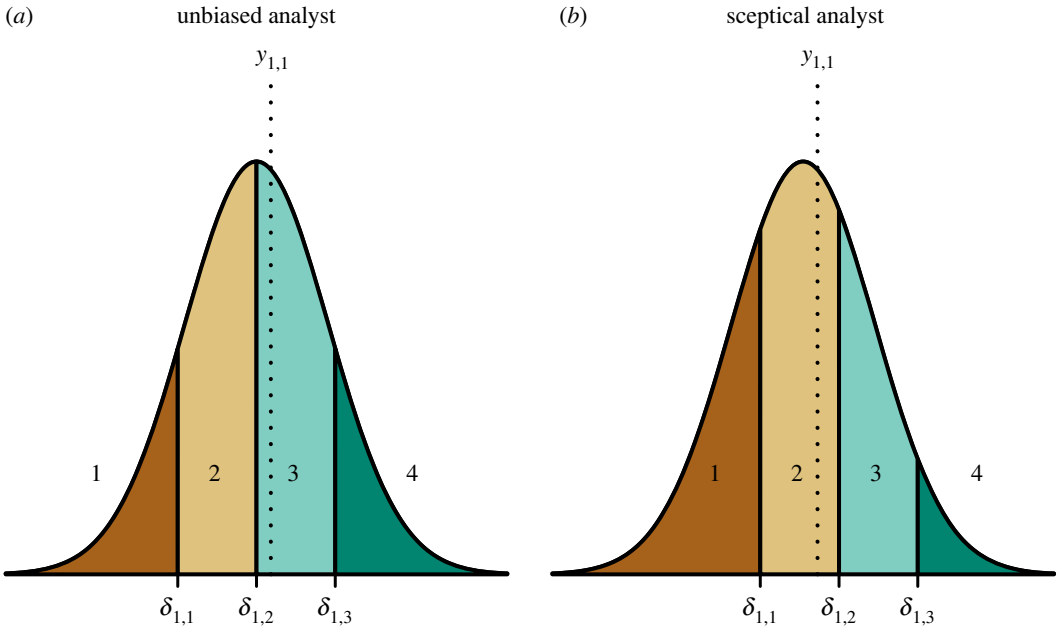

Figure 7. Illustrating the relationship between latent probability distribution over the predicted Likert scores, item appraisal, and decision criteria. In a hypothetical scenario, two analysts with the same item appraisal $y_{r,i}$ differ in their individual decision criteria $\delta_{r,c}$, influenced by a shift parameter. For this particular example, the unbiased analyst ($\beta = 0$; shown in *a*) would respond 'yes, mostly', while the sceptical analyst ($\beta > 0.5$; shown in *b*) would respond 'no, mostly not'.

sample for their item appraisal $y_{i,r}$. The analysts' responses are then determined by where the latent appraisal $y_{i,r}$ falls in relation to their decision criteria $\delta_{c,r}$. The analyst responds with the next higher response category if a latent appraisal for a particular item exceeds a decision criterion, as illustrated in figure 7 and explained mathematically below:

$$x_{r,i} = \begin{cases} 1 & \text{if } y_{r,i} \leq \delta_{r,1} \\ c = 2 \text{ or } 3 & \text{if } \delta_{r,c-1} < y_{r,i} < \delta_{r,c} \\ 4 & \text{if } \delta_{r,3} < y_{r,i}. \end{cases}$$

B.2. Latent consensus

The appraisal of an item $y_{r,i}$ comprises two latent components: the latent consensus (θ_i) and the item difficulty (κ_i). Within the context of a many-analysts project, the primary focus is on estimating the latent consensus as it represents the items' true location on an assumed underlying unidimensional scale. The assumption is that, depending on the item difficulty, the latent consensus is reflected by different analysts with varying degrees of accuracy. In other words, items that elicit polarizing responses from the analysts (i.e. items with lower inter-rater reliability) will be estimated with lower accuracy compared to items that the majority of analysts agree on.

In a scenario where we can estimate the latent consensus without any confounding factors, the item difficulty would be $\kappa_i = 1$, implying that all analysts have the identical information necessary to respond to the item.

B.3. Response bias

As with the item appraisal, the latent decision criteria $\delta_{r,c}$ for each analyst likewise consist of two components: a shift parameter β_r and the response thresholds λ_c:

$$\delta_{r,c} = \lambda_c + \beta_r.$$

The shift parameter determines an analyst's tendency to select lower or higher values on the response scale (i.e. representing individual scepticism). The response thresholds, unaffected by biases, are determined solely by the number of response categories and are identical across all items and analysts, that is, $\lambda_c = \text{logit}(c/C)$. In the scenario of an entirely unbiased analyst, their shift parameter

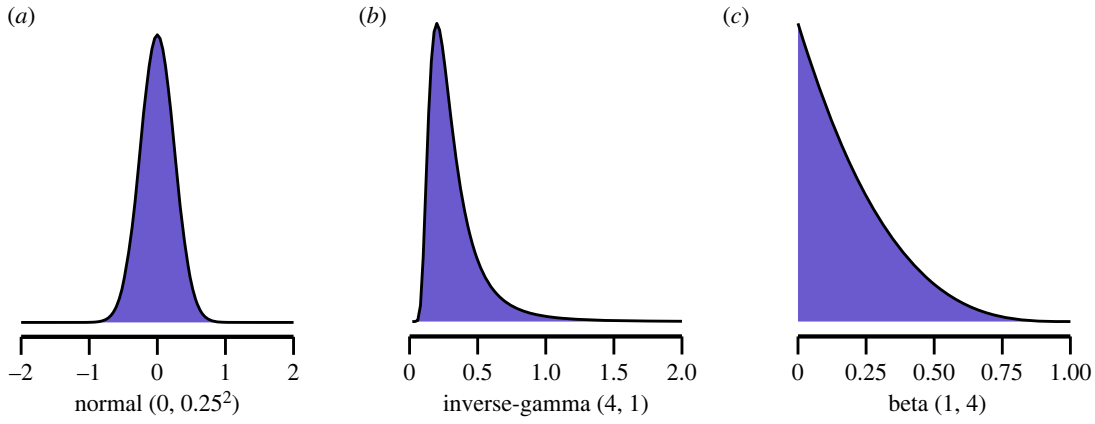

Figure 8. Visualization of the group-level mean prior distributions (*a*), the group-level standard deviation prior distributions (*b*) and the prior distribution on the applicability probabilities (*c*) used in the SEES model.

would be $\beta_r = 0$, suggesting a neutral inclination toward selecting values on the response scale (see figure 7*a*). For shift parameters greater than zero, the individual decision criteria for each item move to the right, leading to lower values on the response scale and consequently more sceptical responses. To illustrate this shift, see figure 7*b* which depicts the latent probability distribution across predicted Likert scores for a sceptical analyst with a shift parameter of $\beta = 0.5$. Conversely, when $\beta_r < 0$, the decision criteria shift to the left, leading to more positive responses.

Bringing together all components of the model, the probabilities of selecting a specific response category can be modelled using the cumulative distribution of the standard logistic distribution, given by $F(x) = (1 + e^{-x})^{-1}$. Given the latent appraisal $y_{r,i}$ and the latent decision criteria δ_r, the responses $x_{r,i}$ then follow an ordered logistic distribution:

$$P(x_{r,i} \mid y_{r,i}, \delta_r) = \begin{cases} 1 - F(y_{r,i} - \delta_{r,1}) & \text{if } x_{r,i} = 1 \\ F(y_{r,i} - \delta_{r,c-1}) - F(y_{r,i} - \delta_{r,c}) & \text{if } 1 < x_{r,i} < 4 \\ F(y_{r,i} - \delta_{r,3}) & \text{if } x_{r,i} = 4. \end{cases}$$

B.4. Prior distributions

We based our prior distributions on the suggestions provided in [76] as a starting point. Subsequently, we refined the values to achieve prior predictions that reflect reasonable response patterns (i.e. an approximately uniform distribution of the predicted responses) but are still vague enough to ensure proper updating of the parameters in light of the data. A visualization of the prior distributions for the group-level means and group-level standard deviations and the prior distribution on the applicability probability are visualized in figure 8.

The applicability probability π_i for each item is assumed to be drawn from a beta distribution that mildly favours items being considered appropriate. The remaining parameters are assumed to be drawn from normal distributions. Due to identifiability constraints discussed in [76], the group-level mean for the item difficulty κ is fixed to 1:

$$\begin{aligned} \pi_i &\sim \text{beta } (1, 4) \\ \theta_i &\sim \text{normal } (\mu_\theta, \sigma_\theta^2) \\ \kappa_i &\sim \text{normal } (1, \sigma_\kappa^2) \\ \text{and} \qquad \beta_r &\sim \text{normal } (\mu_\beta, \sigma_\beta^2). \end{aligned}$$

The group-level means for the consensus and the shift parameter are chosen to be relatively uninformative. Their values are drawn from a normal distribution centred at 0 with standard deviations that favour values centred around zero:

$$\mu_\theta \sim \text{normal } (0, 0.25^2)$$

and

$$\mu_\beta \sim \text{normal } (0, 0.25^2).$$

The standard deviations are drawn from inverse-gamma distributions which allow for moderate heterogeneity for the consensus and individual biases:

$$\sigma_\theta \sim \text{inverse-gamma}\ (4, 1)$$
$$\sigma_\kappa \sim \text{inverse-gamma}\ (4, 1)$$
and
$$\sigma_\beta \sim \text{inverse-gamma}\ (4, 1).$$

B.5. Assumptions

The model is based on several key assumptions. First, it assumes that the probabilities of items being deemed non-applicable are independent across items. For instance, in the subjective evidence subscale, an analyst may feel insufficiently informed to respond to item 5 (*If applicable, does the hypothesized effect or relation vary between subgroups or data exclusion criteria?'*) and consequently select the response option 'not applicable/I do not know'. However, this response may not affect whether they answer 'not applicable/I do not know' to item 3 (*Does your analysis based on the observed data provide substantial evidence for the hypothesized effect or relation?*). A violation of this assumption may overestimate the heterogeneity of the estimated applicability probabilities due to the absence of hierarchical shrinkage. Note that this independence assumption is only made for 'not applicable/I do not know' answer options. For the remaining answer options items and participants are assumed to be related through the hierarchical structure of the model parameters.

Second, the original version of the latent truth rater model included two additional parameters: a parameter describing an analyst's inclination towards extreme responses (i.e. answering more confidently) and a parameter describing the conformity of an analyst to the group opinion. The conformity parameter describes the extent to which an analyst deviates from the group opinion—perhaps due to adopting an unconventional analytic approach. In other words, non-conforming analysts are more prone to give responses that differ from what we would anticipate based solely on the consensus. The current model assumes that analysts have no bias towards extreme responses and that there is a perfect alignment of opinions among analysts. The reason for these assumptions lies in the nature of the SEES, which by design, produces scarce data (with only 8 and 10 data points per analyst for each subscale, respectively) limiting its capacity to capture parameters characterizing each individual analyst. By placing additional assumptions on the extremity bias and group conformity, we were able to reduce the number of parameters and improve the model's ability to recover true parameter values in a scarce data environment.

Third, following the recommendations in [76], the model assumes that an analyst's decision criteria can be effectively described using only the shift parameter, eliminating the need to estimate each threshold separately. Lastly, the model places a sum-to-one constraint on the response thresholds λ_c so that for an unbiased analyst the model predicts *a priori* a uniform distribution over the survey responses. Assumptions three and four resolve identification issues within the model.

B.6. Model validation

To ensure the model aligns with theoretical expectations, we generated prior predictive plots for a sample of $n = 42$ which matches our pilot study's sample size. In the absence of information regarding the evidence supporting a hypothesis, the methodological appropriateness of the research design, or the analysts responding to the SEES, we would expect a uniform distribution of responses. That is, we would expect *a priori* that each response category gets selected equally often. The prior predictive distributions confirm this expectation. Figure 9 presents the prior predictions for both subscales across a hypothetical group of 42 analysts. The response categories 1 to 4 are distributed approximately equally for each item, while the 'not applicable/I do not know' response category is anticipated to be chosen slightly less frequently.

The posterior predictive distribution can be interpreted as the model's attempt to re-describe the behavioural data and constitutes another step in the model validation process. Predictions from an adequate model should resemble the behavioural data. In figure 10, we visualize for each item the relative proportion of observed responses from our pilot study (purple bars) and the model's predicted responses (black dots and 95% credible intervals). Indeed, the posterior model predictions are able to re-describe the observed data.

In addition to examining prior and posterior predictive distributions, we ensured that the proposed computational model as well as the implemented Markov chain Monte Carlo (MCMC) algorithm is able to recover the prior distribution when no data are observed, that the implemented MCMC algorithm returns unbiased estimates, and that the data effectively update the prior beliefs. These additional checks were proposed in [81], based on the recommendations in [82,83]. We conducted these checks for samples

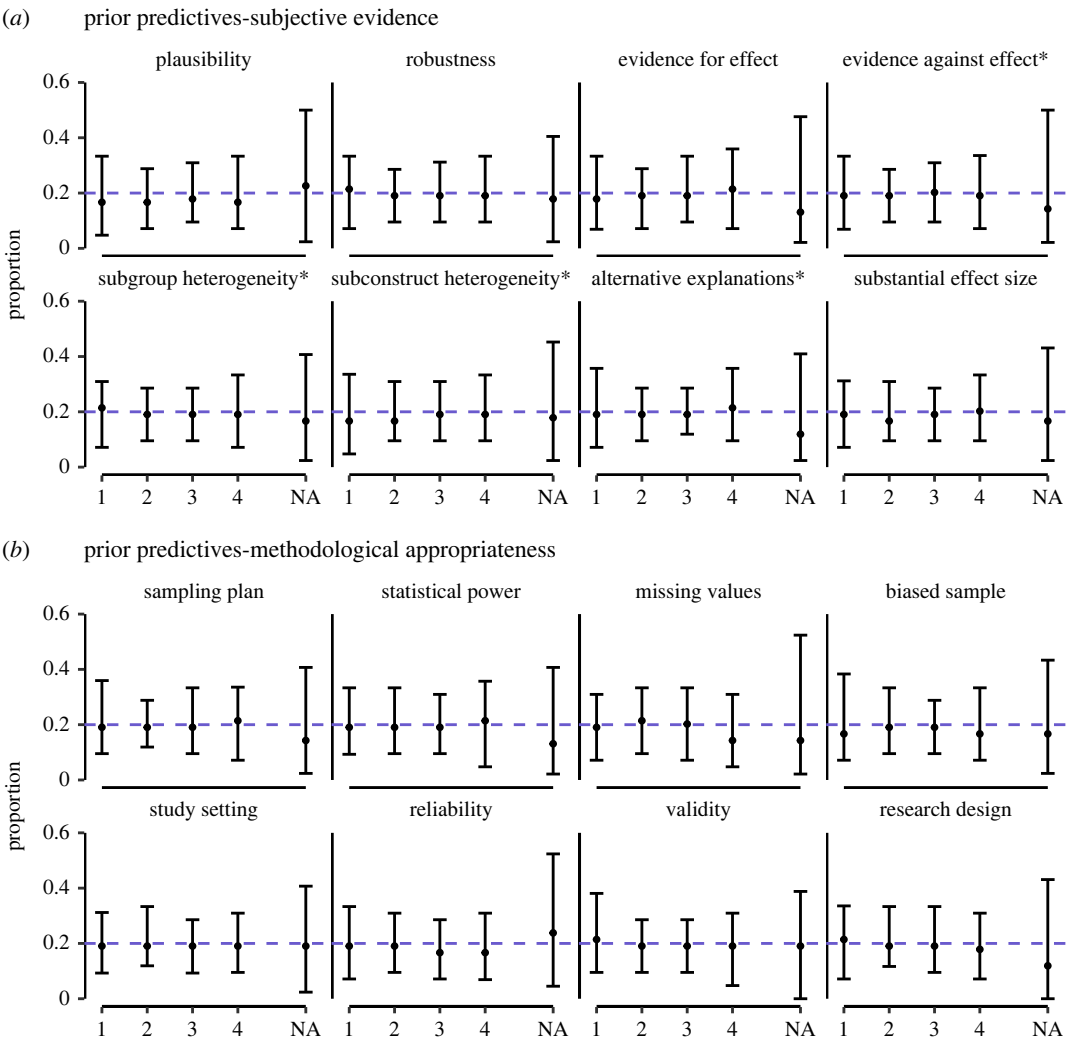

Figure 9. Prior model predictions of data for both subscales. Before having been in contact with the data, the model predicts analysts will select each response category nearly equally often, showing a slightly lower tendency for the 'not applicable/I do not know' response category.

of size $n = 42$ and $n = 20$ with satisfactory model performance. The full results can be accessed in the electronic supplementary material.

Appendix C. Additional results

C.1. Descriptive results

Here, we report the findings of the example data from the Many-Analysts Religion Project using descriptive results, as an alternative to the modelling approach outlined in the main text.

C.1.1. Subjective evidence

The mean of the overall consensus rating for the *subjective evidence* subscale is 3.08; the 95% confidence interval is [2.98, 3.19]. Figure 13*a* shows the correlation between the observed item means and estimated consensus per item, indicating high correspondence between the observed and estimated item metrics ($\rho = 0.89$, 95% CI [0.84, 0.98]).

C.1.2. Methodological appropriateness

The mean of the overall consensus rating for the *methodological appropriateness* subscale is 3.21; the 95% confidence interval is [3.12, 3.3]. Figure 13*b* shows the correlation between the observed item means and estimated consensus per item, again indicating high correspondence between the observed and estimated item metrics ($\rho = 0.86$, 95% CI [0.68, 0.95]).

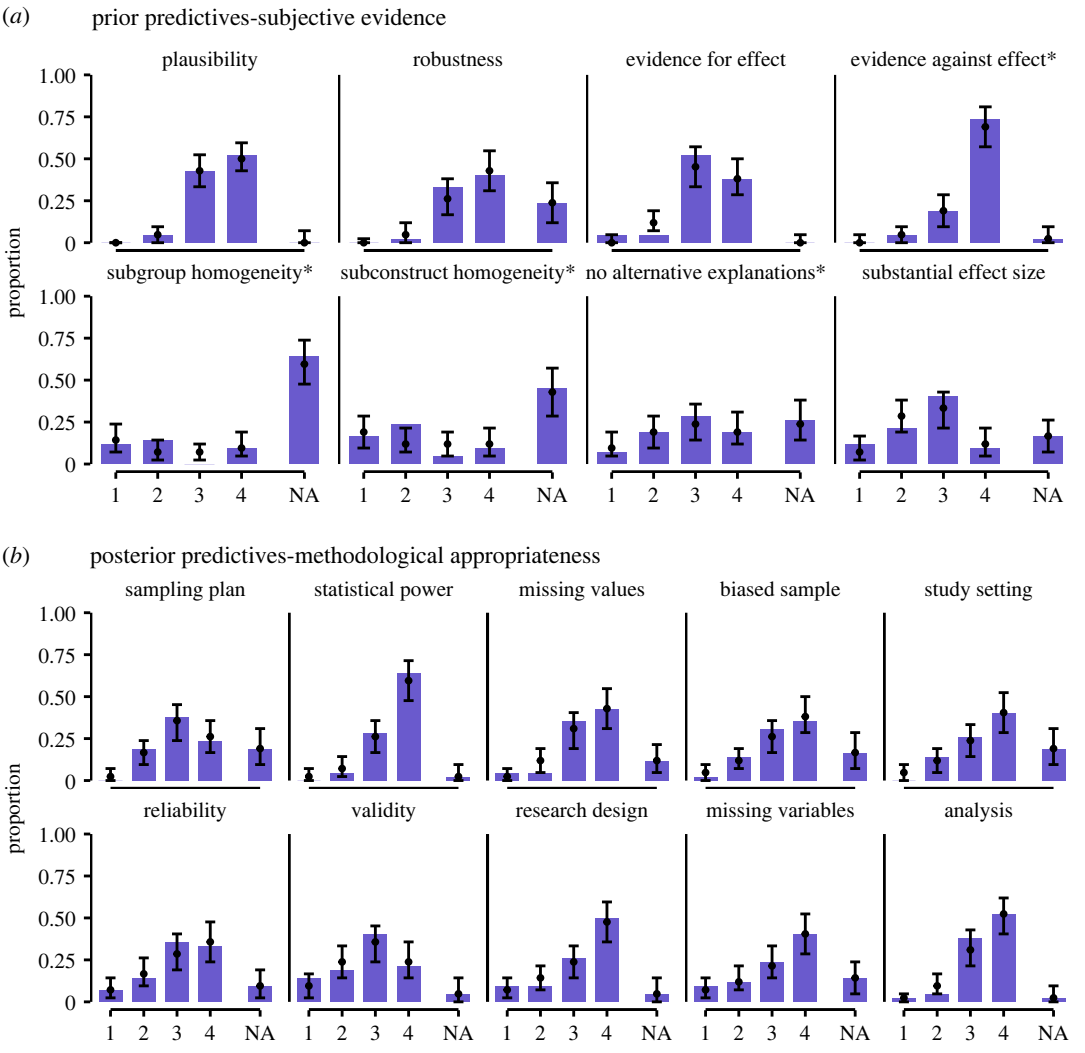

Figure 10. Posterior model predictions of data for both subscales. For each item, the purple bars reflect the observed relative proportion of responses and the black dots plus 95% credible interval reflect the predicted responses from the model.

C.2. Confirmatory factor analysis

To validate the two-component structure of the SEES, we conducted Bayesian confirmatory factor analysis for ordinal data using the `blavaan` package in R [79]. Specifically, we fitted the measurement model assuming the eight *subjective evidence* items to load on one factor and the ten *methodological concerns* items to load on a second factor (covariance between factors: $\rho = 0.47$). All standardized factor loading was above 0.5 except for the 'subgroup homogeneity' (0.35) and 'subconstruct homogeneity' (0.17) items of the subjective evidence scale.

Following recommendations by [84,85], we applied model comparison using the ratio of differences in the leave-one-out cross-validation metric and its standard error to interpret evidence for the two-factor model over the null model (i.e. the independence baseline model). In particular, we found that the two-factor model fitted the data better than the baseline model, with a magnitude of 4.87 SE in LOO differences, which can be considered substantial [85].

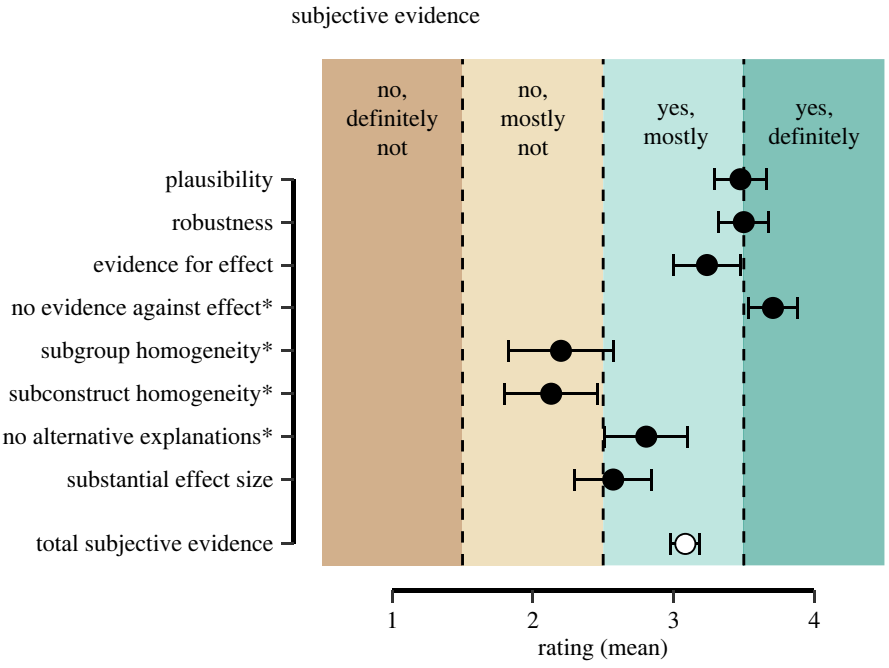

Figure 11. Observed item ratings for the subjective evidence subscale. The black points show the means (plus 95% confidence interval) of the item ratings, including the category thresholds. Items followed by an asterisk reflect items that were reverse-coded. The white point at the bottom reflects the overall mean assessment (plus 95% confidence interval) of the subjective evidence subscale.

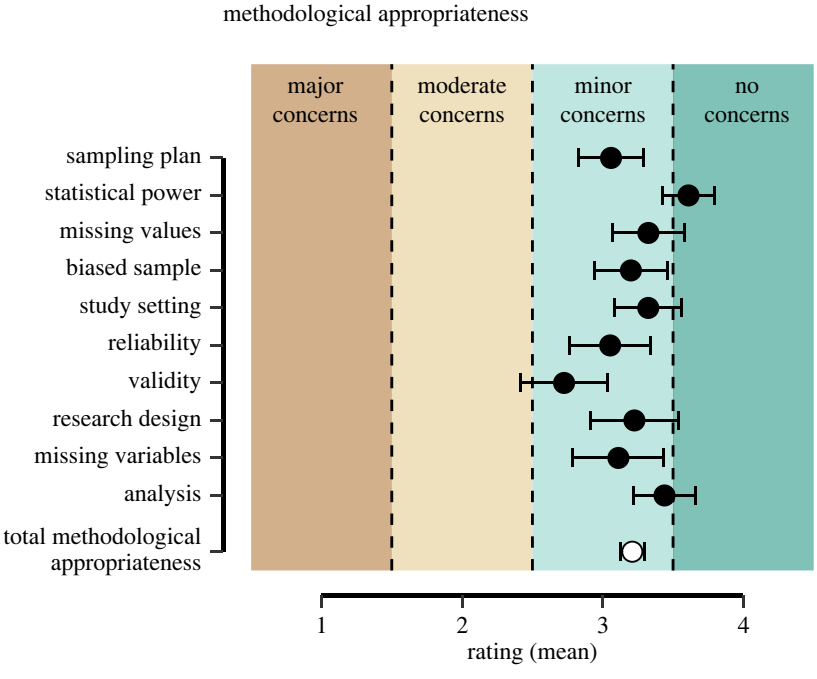

Figure 12. Observed item ratings for the methodological appropriateness subscale. The black points show the means (plus 95% confidence interval) of the item ratings, including the category thresholds. The white point at the bottom reflects the overall mean assessment (plus 95% confidence interval) of the methodological appropriateness subscale.

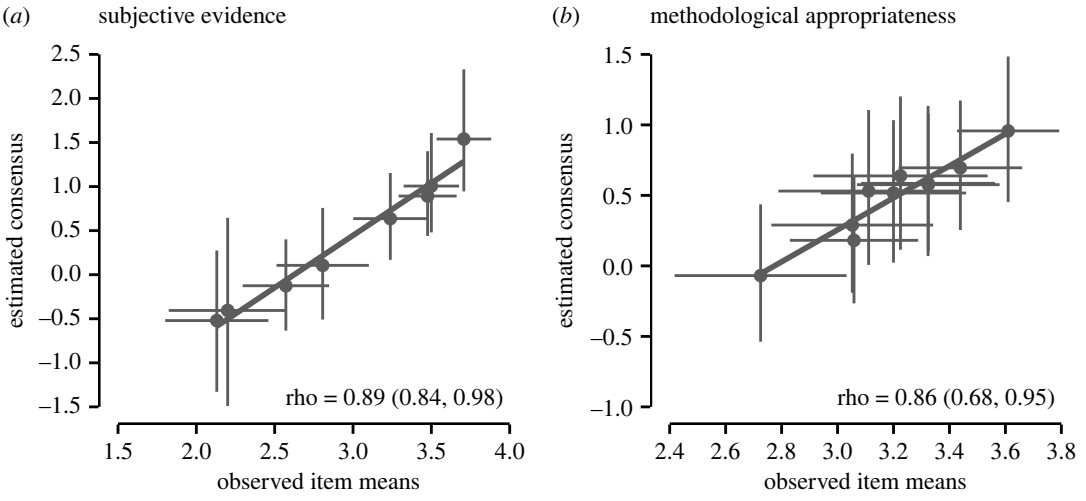

Figure 13. Correlation between the estimated consensus and the observed item means for the subjective evidence subscale (*a*) and the methodological appropriateness subscale (*b*). Vertical error bars reflect the 95% credible interval of the estimated consensus and horizontal error bars reflect the 95% confidence interval of the observed scores. 'rho' gives the Bayesian Spearman correlation between the estimated consensus and observed means.

References

1. Wicherts JM, Veldkamp CL, Augusteijn HE, Bakker M, Van Aert R, Van Assen MA. 2016 Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid *p*-hacking. *Front. Psychol.* **7**, 1832. (doi:10.3389/fpsyg.2016.01832)
2. Gelman A, Loken E. 2014 The statistical crisis in science data-dependent analysis—a 'Garden of Forking Paths'—explains why many statistically significant comparisons don't hold up. *Am. Sci.* **102**, 460. (doi:10.1511/2014.111.460)
3. Holzmeister F, Johannesson M, Böhm R, Dreber A, Huber J, Kirchler M. 2024 Heterogeneity in effect size estimates: empirical evidence and practical implications. *MetaArXiv*. (doi:10.31222/osf.io/583un)
4. Wagenmakers EJ, Sarafoglou A, Aczel B. 2022 One statistical analysis must not rule them all. *Nature* **605**, 423–425. (doi:10.1038/d41586-022-01332-8)
5. Wagenmakers EJ, Sarafoglou A, Aczel B. 2023 Facing the unknown unknowns of data analysis. *Curr. Dir. Psychol. Sci.* **32**, 362–368. (doi:10.1177/09637214231168565)
6. Silberzahn R, Uhlmann EL. 2015 Crowdsourced research: many hands make tight work. *Nature* **526**, 189–191. (doi:10.1038/526189a)
7. Aczel B *et al.* 2021 Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife* **10**, e72185. (doi:10.7554/eLife.72185)
8. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. 2016 Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.* **11**, 702–712. (doi:10.1177/1745691616658637)
9. Patel CJ, Burford B, Ioannidis JPA. 2015 Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J. Clin. Epidemiol.* **68**, 1046–1058. (doi:10.1016/j.jclinepi.2015.05.029)
10. Odenbaugh J, Alexandrova A. 2011 Buyer beware: robustness analyses in economics and biology. *Biol. Phil.* **26**, 757–771. (doi:10.1007/s10539-011-9278-y)
11. Hoogeveen S, Berkhout SW, Gronau QF, Wagenmakers EJ, Haaf JM. 2023 Improving statistical analysis in team science: the case of a Bayesian multiverse of Many Labs 4. *Adv. Methods Pract. Psychol. Sci.* **6**, 25152459231182318. (doi:10.1177/25152459231182318)
12. Modecki KL, Low-Choy S, Uink BN, Vernon L, Correia H, Andrews K. 2020 Tuning into the real effect of smartphone use on parenting: a multiverse analysis. *J. Child Psychol. Psychiatry* **61**, 855–865. (doi:10.1111/jcpp.13282)
13. Donnelly S, Brooks PJ, Homer BD. 2019 Is there a bilingual advantage on interference-control tasks? A multiverse meta-analysis of global reaction time and interference cost. *Psychon. Bull. Rev.* **26**, 1122–1147. (doi:10.3758/s13423-019-01567-z)
14. Palpacuer C, Hammas K, Duprez R, Laviolle B, Ioannidis JPA, Naudet F. 2019 Vibration of effects from diverse inclusion/exclusion criteria and analytical choices: 9216 different ways to perform an indirect comparison meta-analysis. *BMC Med.* **17**, 174. (doi:10.1186/s12916-019-1409-3)
15. Klau S, Hoffmann S, Patel CJ, Ioannidis JP, Boulesteix AL. 2021 Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework. *Int. J. Epidemiol.* **50**, 266–278. (doi:10.1093/ije/dyaa164)
16. Botvinik-Nezer R *et al.* 2020 Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88. (doi:10.1038/s41586-020-2314-9)
17. Trübutschek D *et al.* 2023 EEGManyPipelines: a large-scale, grassroots multi-analyst study of electroencephalography analysis practices in the wild. *J. Cogn. Neurosci.* **36**, 217–224. (doi:10.1162/jocn_a_02087)
18. Breznau N *et al.* 2022 Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proc. Natl Acad. Sci. USA* **119**, e2203150119. (doi:10.1073/pnas.2203150119)
19. Boehm U *et al.* 2018 Estimating across-trial variability parameters of the diffusion decision model: expert advice and recommendations. *J. Math. Psychol.* **87**, 46–75. (doi:10.1016/j.jmp.2018.09.004)
20. van Dongen NNN *et al.* 2019 Multiple perspectives on inference for two simple statistical scenarios. *Am. Stat.* **73**, 328–339. (doi:10.1080/00031305.2019.1565553)
21. Silberzahn R *et al.* 2018 Many analysts, one data set: making transparent how variations in analytic choices affect results. *Adv. Methods Pract. Psychol. Sci.* **1**, 337–356. (doi:10.1177/2515245917747646)
22. Bastiaansen JA *et al.* 2020 Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *J. Psychosom. Res.* **137**, 110211. (doi:10.1016/j.jpsychores.2020.110211)
23. Hoogeveen S *et al.* 2023 A many-analysts approach to the relation between religiosity and well-being. *Religion Brain Behav.* **13**, 237–283. (doi:10.1080/2153599X.2022.2070255)
24. Fillard P *et al.* 2011 Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. *NeuroImage* **56**, 220–234. (doi:10.1016/j.neuroimage.2011.01.032)

25. Maier-Hein KH *et al.* 2017 The challenge of mapping the human connectome based on diffusion tractography. *Nat. Commun.* **8**, 1349. (doi:10.1038/s41467-017-01285-x)
26. Veronese M *et al.* 2021 Reproducibility of findings in modern PET neuroimaging: insight from the NRM2018 grand challenge. *J. Cereb. Blood Flow Metab.* **41**, 2778–2796. (doi:10.1177/0271678X211015101)
27. Huntington-Klein N *et al.* 2021 The influence of hidden researcher decisions in applied microeconomics. *Econ. Inq.* **59**, 944–960. (doi:10.1111/ecin.12992)
28. Menkveld AJ *et al.* 2021 Non-standard errors. *J. Finance* **79**, 2339–2390. (doi:10.1111/jofi.13337)
29. Scientific Pandemic Influenza Group on Modelling. 2020 SPI-M-O: Consensus Statement on COVID-19, 8 October 2020.
30. Gould E *et al.* 2023 Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology. *EcoEvoRxiv*. (doi:10.32942/X2GG62)
31. Oza A. 2023 Reproducibility trial: 246 biologists get different results from same data sets. *Nature* **622**, 677–678. (doi:10.1038/d41586-023-03177-1)
32. Dutilh G *et al.* 2019 The quality of response time data inference: a blinded, collaborative assessment of the validity of cognitive models. *Psychon. Bull. Rev.* **26**, 1051–1069. (doi:10.3758/s13423-017-1417-2)
33. Starns JJ *et al.* 2019 Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. *Adv. Meth. Pract. Psychol. Sci.* **2**, 335–349. (doi:10.1177/2515245919869583)
34. Schweinsberg M *et al.* 2021 Same data, different conclusions: radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organ. Behav. Hum. Decis. Process.* **165**, 228–249. (doi:10.1016/j.obhdp.2021.02.003)
35. Salganik MJ *et al.* 2020 Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc. Natl Acad. Sci. USA* **117**, 8398–8403. (doi:10.1073/pnas.1915006117)
36. Kümpel H, Hoffmann S. 2022 A formal framework for generalized reporting methods in parametric settings (https://arxiv.org/abs/2211.02621)
37. Coretta S *et al.* 2023 Multidimensional signals and analytic flexibility: estimating degrees of freedom in human-speech analyses. *Adv. Methods Pract. Psychol. Sci.* **6**, 1–29. (doi:10.1177/25152459231162567)
38. McNamara AA. 2023 The impact (or lack thereof) of analysis choice on conclusions with likert data from the many analysts religion project. *Religion Brain Behav.* **13**, 324–326. (doi:10.1080/2153599X.2022.2070256)
39. van Lissa CJ. 2023 Complementing preregistered confirmatory analyses with rigorous, reproducible exploration using machine learning. *Religion Brain Behav.* **13**, 347–351. (doi:10.1080/2153599X.2022.2070254)
40. Hanel PHP, Zarzeczna N. 2023 From multiverse analysis to multiverse operationalisations: 262,143 ways of measuring well-being. *Religion Brain Behav.* **13**, 309–313. (doi:10.1080/2153599X.2022.2070259)
41. Krypotos AM, Klein R, Jong J. 2023 Resolving religious debates through a multiverse approach. *Religion Brain Behav.* **13**, 318–320. (doi:10.1080/2153599X.2022.2070261)
42. Murphy J, Martinez N. 2023 Quantifying religiosity: a comparison of approaches based on categorical self-identification and multidimensional measures of religious activity. *Religion Brain Behav.* **13**, 327–329. (doi:10.1080/2153599X.2022.2070252)
43. Smith E. 2023 Individual-level versus country-level moderation. *Religion Brain Behav.* **13**, 342–344. (doi:10.1080/2153599X.2022.2070246)
44. Atkinson QD, Claessens S, Fischer K, Forsyth GL, Kyritsis T, Wiebels K, Moreau D. 2023 Being specific about generalisability. *Religion Brain Behav.* **13**, 284–286. (doi:10.1080/2153599X.2022.2070251)
45. Vogel V, Prenoveau J, Kelchtermans S, Magyar-Russell G, McMahon C, Ingendahl M, Schaumans CBC. 2023 Different facets, different results: the importance of considering the multidimensionality of constructs. *Religion Brain Behav.* **13**, 351–356. (doi:10.1080/2153599X.2022.2070262)
46. Pearson HI, Lo RF, Sasaki JY. 2023 How do culture and religion interact worldwide? A cultural match approach to understanding religiosity and well-being in the Many Analysts Religion Project. *Religion Brain Behav.* **13**, 329–336. (doi:10.1080/2153599X.2022.2070265)
47. Schreiner MR, Mercier B, Frick S, Wiwad D, Schmitt MC, Kelly JM, Quevedo Pütter J. 2023 Measurement issues in the many analysts religion project. *Religion Brain Behav.* **13**, 339–341. (doi:10.1080/2153599X.2022.2070260)
48. Ross RM, Sulik J, Buczny J, Schivinski B. 2023 Many analysts and few incentives. *Religion Brain Behav.* **13**, 336–339. (doi:10.1080/2153599X.2022.2070248)
49. Edelsbrunner PA, Sebben S, Frisch LK, Schüttengruber V, Protzko J, Thurn CM. 2023 How to understand a research question—a challenging first step in setting up a statistical model. *Religion Brain Behav.* **13**, 306–309. (doi:10.1080/2153599X.2022.2070258)
50. Hoogeveen S, Sarafoglou A, van Elk M, Wagenmakers EJ. 2023 Many-analysts religion project: reflection and conclusion. *Religion Brain Behav.* **13**, 356–363. (doi:10.1080/2153599X.2022.2070263)
51. Mathur MB, Covington C, VanderWeele TJ. 2023 Variation across analysts in statistical significance, yet consistently small effect sizes. *Proc. Natl. Acad. Sci. USA* **120**, e2218957120. (doi:10.1073/pnas.2218957120)
52. Young C, Holsteen K. 2017 Model uncertainty and robustness: a computational framework for multimodel analysis. *Sociol. Methods Res.* **46**, 3–40. (doi:10.1177/0049124115610347)
53. Critical Appraisal Skills Programme. 2018 CASP cohort study checklist. Technical report.
54. Critical Appraisal Skills Programme. 2018 CASP qualitative checklist. Technical report.
55. Briner RB, Denyer D. 2012 Systematic review and evidence synthesis as a practice and scholarship tool. In *The Oxford handbook of evidence-based management*, Oxford Library of Psychology (ed. DM Rousseau), pp. 112–129. Oxford, UK: Oxford Academic. (doi:10.1093/oxfordhb/9780199763986.013.0007)
56. Lewin S *et al.* 2018 Applying GRADE-CERQual to qualitative evidence synthesis findings—Paper 2: how to make an overall CERQual assessment of confidence and create a summary of qualitative findings table. *Implement. Sci.* **13**, 10. (doi:10.1186/s13012-017-0689-2)
57. Spencer L, Ritchie J, Lewis J, Dillon L. 2004 *Quality in qualitative evaluation: a framework for assessing research evidence*. Technical report. London, UK: Government Chief Social Researcher's Office.
58. McKenna HP. 1994 The Delphi technique: a worthwhile research approach for nursing? *J. Adv. Nurs.* **19**, 1221–1225. (doi:10.1111/j.1365-2648.1994.tb01207.x)
59. Aczel B *et al.* 2020 A consensus-based transparency checklist. *Nat. Hum. Behav.* **4**, 4–6. (doi:10.1038/s41562-019-0772-6)
60. Colvin CJ *et al.* 2018 Applying GRADE-CERQual to qualitative evidence synthesis findings—Paper 4: how to assess coherence. *Implement. Sci.* **13**, 13. (doi:10.1186/s13012-017-0691-8)
61. Krosnick JA. 2010 Questionnaire design. In *Handbook of survey research* (eds P Rossi, J Wright, A Anderson), pp. 263–313. Bingley, UK: Emerald Group Publishing.
62. Dawes J. 2008 Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *Int. J. Market Res.* **50**, 61–104. (doi:10.1177/147078530805000106)
63. Revilla MA, Saris WE, Krosnick JA. 2014 Choosing the number of categories in agree–disagree scales. *Sociol. Methods Res.* **43**, 73–97. (doi:10.1177/0049124113509605)
64. Alwin DF, Krosnick JA. 1991 The reliability of survey attitude measurement: the influence of question and respondent attributes. *Sociol. Methods Res.* **20**, 139–181. (doi:10.1177/0049124191020001005)
65. Raaijmakers QA, Van Hoof A, Verbogt TF, Vollebergh WA. 2000 Adolescents' midpoint responses on Likert-type scale items: neutral or missing values? *Int. J. Public Opin. Res.* **12**, 208–216. (doi:10.1093/ijpor/12.2.209)
66. Chyung SY, Roberts K, Swanson I, Hankinson A. 2017 Evidence-based survey design: the use of a midpoint on the Likert scale. *Perform. Improv.* **56**, 15–23. (doi:10.1002/pfi.21727)

67. Nadler JT, Weston R, Voyles EC. 2015 Stuck in the middle: the use and interpretation of mid-points in items on questionnaires. *J. General Psychol.* **142**, 71–89. (doi:10.1080/00221309.2014.994590)
68. Romney AK, Weller SC, Batchelder WH. 1986 Culture as consensus: a theory of culture and informant accuracy. *Am. Anthropol.* **88**, 313–338. (doi:10.1525/aa.1986.88.2.02a00020)
69. Anders R, Batchelder WH. 2015 Cultural consensus theory for the ordinal data case. *Psychometrika* **80**, 151–181. (doi:10.1007/s11336-013-9382-9)
70. Anders R, Batchelder WH. 2012 Cultural consensus theory for multiple consensus truths. *J. Math. Psychol.* **56**, 452–469. (doi:10.1016/j.jmp.2013.01.004)
71. Batchelder WH, Anders R. 2012 Cultural consensus theory: comparing different concepts of cultural truth. *J. Math. Psychol.* **56**, 316–332. (doi:10.1016/j.jmp.2012.06.002)
72. Samejima F. 1969 Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monogr. Suppl.* **34**, 100.
73. Takane Y, de Leeuw J. 1987 On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika* **52**, 393–408. (doi:10.1007/BF02294363)
74. Oravecz Z, Vandekerckhove J, Batchelder WH. 2014 Bayesian cultural consensus theory. *Field Methods* **26**, 207–222. (doi:10.1177/1525822X13520280)
75. Oravecz Z, Anders R, Batchelder WH. 2015 Hierarchical Bayesian modeling for test theory without an answer key. *Psychometrika* **80**, 341–364. (doi:10.1007/s11336-013-9379-4)
76. van den Bergh D, Bogaerts S, Spreen M, Flohr R, Vandekerckhove J, Batchelder WH, Wagenmakers EJ. 2020 Cultural consensus theory for the evaluation of patients' mental health scores in forensic psychiatric hospitals. *J. Math. Psychol.* **98**, 102383. (doi:10.1016/j.jmp.2020.102383)
77. Carpenter B *et al.* 2017 Stan: a probabilistic programming language. *J. Stat. Softw.* **76**, 1–32. (doi:10.18637/jss.v076.i01)
78. Hoffman MD, Gelman A. 2014 The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593–1623.
79. Merkle EC, Rosseel Y. 2018 Blavaan: Bayesian structural equation models via parameter expansion. *J. Stat. Softw.* **85**, 1–30. (doi:10.18637/jss.v085.i04)
80. Sarafoglou A *et al.* 2024 Subjective evidence evaluation survey for many-analysts studies. OSF. (https://osf.io/jk674/)
81. Sarafoglou A *et al.* 2024 Subjective evidence evaluation survey for many-analysts studies. Figshare. (doi:10.6084/m9.figshare.c.7360072)
82. Kucharský Š, Tran NH, Veldkamp K, Raijmakers M, Visser I. 2021 Hidden Markov models of evidence accumulation in speeded decision tasks. *Comput. Brain Behav.* **4**, 416–441. (doi:10.1007/s42113-021-00115-0)
83. Talts S, Betancourt M, Simpson D, Vehtari A, Gelman A. 2018 Validating Bayesian inference algorithms with simulation-based calibration (https://arxiv.org/abs/1804.06788)
84. Schad DJ, Betancourt M, Vasishth S. 2021 Toward a principled Bayesian workflow in cognitive science. *Psychol. Methods* **26**, 103–126. (doi:10.1037/met0000275)
85. Merkle EC, Rosseel Y, Goodrich B. 2018 blavaan: Bayesian structural equation models via parameter expansion. *J. Stat. Softw.* **85**, 1–30. (doi:10.18637/jss.v085.i04)
86. Vehtari A, Gelman A, Gabry J. 2017 Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432. (doi:10.1007/s11222-016-9696-4)