Subjective Evidence Evaluation Survey For Multi-Analyst 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[4], 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[4], 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] University of Amsterdam [2] Utrecht University [3] Institute of Psychology, ELTE Eötvös Lorénd University, Budapest, Hungary [4] Heymans Institute for Psychological Research, University of Groningen [5] Allen School of Computer Science & Engineering, University of Washington [6] Hebrew University of Jerusalem [7] Institute for Psychology, University of Münster, Germany [8] Center for Depression, Anxiety and Stress Research, McLean Hospital, Belmont, MA [9] University of Idaho [10] Stockholm School of Economics [11] Department of Psychology, Leiden University [12] Department of Statistics Ludwig-Maximilians-Universität München [13] University of Innsbruck [14] Seattle University [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, USA [16] University of Conneticut [17] University Paris-Saclay [18] Vrije Universiteit Amsterdam [19] Karolinska Institutet [20] ESMT Berlin [21] Department of Statistics LMU Munich [22] University College London [23] University of Illinois - Urbana-Champaign [24] University of Virginia [25] Max Planck Institute for Empirical Aesthetics, Germany [26] University of Leuven, Belgium [27] INSEAD [28] Department of Methodology and Statistics, Tilburg University [29] Dartmouth College [30] Department of Psychiatry, Harvard Medical School, Boston, MA [31] Neurospin CEA [32] The Institute for Medical Information Processing, Biometry, and Epidemiology, LMU Munich

Abstract

Multi-analyst studies explore how well an empirical claim withstands plausible alternative analyses of the same data set 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 data set, 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 multi-analyst study.

Keywords: Open Science, Team Science, Scientific Transparency, Metascience, Many Analysts

Introduction

Researchers adopt a wide range of approaches when analyzing data, and their equally justifiable choices about statistical procedures, data processing, and the inclusions of covariates can affect the conclusions they draw (Gelman & Loken, 2014; Wicherts et al., 2016). 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., Holzmeister et al., 2024), resulting in overconfidence and model myopia (Aczel et al., 2021; Silberzahn & Uhlmann, 2015; Wagenmakers et al., 2022, 2023).

The typical way to assess robustness of an empirical claim is through meta-analysis. However, meta-analyses may suffer from publication bias, their conclusions may differ between meta-analytic techniques, and estimated meta-analytic effect sizes might be as much as three times larger than in preregistered multiple-site replication studies (de Vrieze, 2018; Kvarven et al., 2020; van Elk et al., 2015).

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 (Patel et al., 2015; Steegen et al., 2016) and multi-analyst approaches (Silberzahn and Uhlmann, 2015; but see Odenbaugh and Alexandrova, 2011 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., Donnelly et al., 2019; Hoogeveen, Berkhout, et al., 2023; Klau et al., 2021; Modecki et al., 2020; Palpacuer et al., 2019; Patel et al., 2015). In a multi-analyst project, different independent

Correspondence concerning this article should be addressed to: Alexandra Sarafoglou, Nieuwe Achtergracht 129B, 1001 NK Amsterdam, The Netherlands, E-mail: alexandra.sarafoglou@gmail.com.

analysis teams analyze the same data set (e.g., Bastiaansen et al., 2020; Boehm et al., 2018; Botvinik-Nezer et al., 2020; Breznau et al., 2022; Hoogeveen, Sarafoglou, Aczel, et al., 2023; Silberzahn et al., 2018; Trübutschek et al., 2023; van Dongen et al., 2019). In both cases, the end result is an evaluation of the consistency of the observed outcomes across all analyses.

Multi-analyst 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 multi-analyst 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.

Multi-analyst projects are a recent innovation, but they have already been adopted in many different fields, including neuroscience (Botvinik-Nezer et al., 2020; Fillard et al., 2011; Maier-Hein et al., 2017; Trübutschek et al., 2023; Veronese et al., 2021), economics (Huntington-Klein et al., 2021; Menkveld et al., 2021), epidemiology (Scientific Pandemic Influenza Group on Modelling, 2020), ecology (Gould et al., 2023; Oza, 2023), cognitive science (Boehm et al., 2018; Dutilh et al., 2019; Starns et al., 2019), and psychology (Bastiaansen et al., 2020; Breznau et al., 2022; Hoogeveen, Sarafoglou, Aczel, et al., 2023; Salganik et al., 2020; Schweinsberg et al., 2021; Silberzahn et al., 2018; van Dongen et al., 2019). 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., Menkveld et al., 2021; Schweinsberg et al., 2021, but see Hoogeveen, Sarafoglou, Aczel, et al., 2023).

Beyond Effect Sizes: Acknowledging Insights and Concerns of Analysis Teams

The multi-analyst 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., Silberzahn et al., 2018 or a standardized beta coefficient, e.g., Hoogeveen, Sarafoglou, Aczel, et al., 2023, but see Trübutschek et al., 2023). These effect size estimates are (visually) summarized to provide an overall impression of the results (but see Kümpel and Hoffmann, 2022 and Coretta et al., 2023 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. Odenbaugh & Alexandrova, 2011).

Commentaries on the recently published Many-Analysts Religion Project (Hoogeveen, Sarafoglou, Aczel, et al., 2023), studying the relationship between self-reported well-being and religiosity, challenge all three assumptions. 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 modeling (McNamara, 2023), machine learning (van Lissa, 2023), and even multiverse analyses (Hanel & Zarzeczna, 2023; Krypotos et al., 2023).[1] Second, many teams presented more nuanced interpretations of the primary effect based on sub-group analyses or multivariate approaches (e.g., Atkinson et al., 2023; Murphy & Martinez, 2023; Pearson et al., 2023; Smith, 2023; Vogel et al., 2023) which helped determine the conditions under which the hypothesized relation occurred. Third, some teams raised concerns about measurement invariance in the data themselves (e.g., Ross et al., 2023; Schreiner et al., 2023). Others criticized the formulation of the research question (e.g., Edelsbrunner et al., 2023; Murphy & Martinez, 2023), an issue that surfaced in previous multi-analyst projects (van Dongen

[1] Alternative approaches for synthesizing outcomes in multi-analyst 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 (see e.g., Mathur et al., 2023).

et al., 2019). 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 (Hoogeveen, Sarafoglou, van Elk, et al., 2023).

Assessment of Subjective Evidence

Although measuring the distribution of plausible effect sizes can provide important insights about the robustness of an empirical result (e.g., Coretta et al., 2023; Kümpel & Hoffmann, 2022), we argue that it is incomplete. 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 is uncommon in the quantitative social and behavioral 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 multi-analyst project. Instead, we propose a systematic assessment of each team's subjective evaluation of the evidence, design, and data.

Other fields commonly use the subjective evaluation of 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 (Briner, Denyer, et al., 2012; Critical Appraisal Skills Programme, 2018a, 2018b). 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 (Lewin et al., 2018; Spencer et al., 2004).

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 multi-analyst projects. Multi-analyst 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.

Current Study

The aim of the current project is twofold. First, we aimed to advance multi-analyst 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 multi-analyst 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 multi-analyst 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.

Development of the SEES

The idea for the SEES arose from the experience some of us (AS and SH) had in leading a multi-analyst project (i.e., Hoogeveen, Sarafoglou, Aczel, et al., 2023), 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 (McKenna, 1994) as implemented in Aczel et al. (2021) and Aczel et al. (2020). 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.

Creating the Initial Item List. During the planning phase, authors AS, SH, and EJW drafted an initial item list containing 22 items, which was based on checklist and guidance articles on systematic literature reviews (Briner, Denyer, et al., 2012) and evaluating qualitative evidence (Colvin et al., 2018; Spencer et al., 2004), and items used in previous multi-analyst studies (Hoogeveen, Sarafoglou, Aczel, et al., 2023; Silberzahn et al., 2018).[2]

Recruitment of the Expert Panel. On 25 November 2022, we contacted 93 experts, including project leaders of multi-analyst and multiverse studies listed in Aczel et al. (2021), 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 Briner, Denyer, et al., 2012; Critical Appraisal Skills Programme, 2018a, 2018b; Lewin et al., 2018; Spencer et al., 2004). Furthermore, we invited measurement and general methodology experts, selecting them based on our knowledge of publications on cautionary

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

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, 7 declined our invitation, 38 did not respond to our request, and 3 invitations bounced. Of these 45 experts, 37 finished all three consensus rounds.

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. 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. All items received approval from panel members during the discussion round. A detailed description of the different stages of the Delphi method can be found in https://osf.io/jk674/.

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.

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.

Subjective Evidence Subscale

The subjective evidence subscale consists of 8 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.

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 skeptical to favorable.

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 favors 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 vs. 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.*

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.

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, 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.*
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.*

Computational Model

The computational model we developed to synthesize responses to the SEES is based on cultural consensus theory (Anders & Batchelder, 2012; Anders & Batchelder, 2015; Batchelder & Anders, 2012; Romney et al., 1986). 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 multi-analyst studies, the cultural consensus theory model estimates the analysis teams' collective opinion for each of the scale items, henceforth referred to as *item truths*, as well as overall, on the evidence in the data related to the research question.

Compared to the use of standard descriptive statistics such as sum scores or means, the application of cultural consensus theory brings three main advantages. First, the proposed model is hierarchical which takes into account both the similarities between analysis teams and the nested structure of items within a scale. Second, the model corrects for response biases of analysis teams (i.e., levels of skepticism defined as an analyst's tendency to select lower (vs. higher) values on the response scale) and differences among items (i.e., whether the items elicit polarizing responses from the analysts) in the estimates of the item truths. Third, the model is embedded within a Bayesian framework, facilitating the quantification of uncertainty for the parameters of interest (Oravecz et al., 2015; Oravecz et al., 2014; van den Bergh et al., 2020).

Specifically, the applied computational model is an adapted version of the latent truth rater model proposed in Anders and Batchelder (2015) and extended by van den Bergh et al. (2020). The model is implemented in the Stan programming language using the No-U-Turn sampler (Carpenter et al., 2017; Hoffman & Gelman, 2014). Appendix B contains a formal description of the model.

An Example Application of the SEES

To showcase the intended use of the SEES in a multi-analyst 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 ($M = 4.07$,

Table 1

Descriptive Statistics for the Subjective Evidence Subscale for the Pilot Data

Label	M	SD	No answer
Plausibility	3.48	0.59	0%
Robustness	3.50	0.57	23.8%
Evidence for effect	3.24	0.76	0%
No evidence against effect*	3.71	0.56	2.4%
Subgroup homogeneity*	2.20	1.21	64.3%
Subconstruct homogeneity*	2.13	1.06	45.2%
No alternative explanations*	2.81	0.95	26.2%
Substantial effect size	2.57	0.88	16.7%

Note. All items were answered on a 4-point scale. Items followed by an asterisk * have been reverse-coded and their labels have been changed for interpretability.

SD = 0.64 for responders vs. M = 4.01, SD = 0.71 for non-responders) and substantive knowledge (M = 2.76, SD = 1.41 for responders vs. M = 2.63, SD = 1.22 for non-responders). Nevertheless, these data merely serve as an illustration for how to use and analyze 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 skepticism.[3]

Subjective Evidence

Figure 1 shows the model-based item truths for each item of the subjective evidence subscale, including the average response category thresholds and their labels. The item truths represent the true location of the items on an assumed underlying unidimensional

[3] Note that the Many-Analysts Religion Project analysis teams filled out a preliminary version of the SEES in which the subjective evidence subscale were 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'.

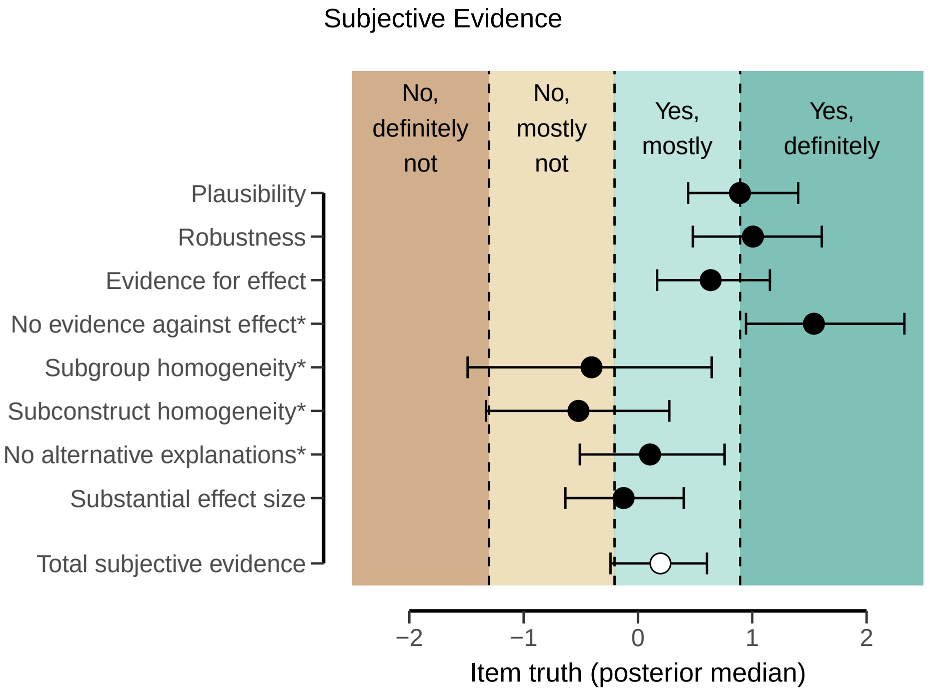


Figure 1

*Estimated item truths for the subjective evidence subscale. The black points show the posterior medians (plus 95% credible interval) of the item truths, 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.*

scale ranging from minus to plus infinity. When interpreting the item truths, it is crucial to note that they can only be interpreted in relation to the response category thresholds, which are represented by different colors in the figure. The posterior median and 95% credible interval of the overall item truth 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.'). For item 5 ('subgroup homogeneity'), the analysis teams seem neutral regarding whether effects differed across subgroup analyses. The large credible interval for this item reflects the uncertainty in the estimated item truth, as more than half of the analysis teams considered the item not applicable (suggesting that they did not conduct subgroup analyses). Figure 2A shows the correlation between the observed item means and estimated item truths, indicating high correspondence between the observed and estimated item metrics.

Methodological Appropriateness

Figure 3 shows the model-based item truths for each item of the 10 items of the methodological appropriateness subscale. Compared to the subjective evidence subscale, the latent item truths for the methodological appropriateness subscale appear more similar. The posterior median of the overall item truth 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). The correlation between the observed item means and estimated item truths shown in Figure 2B again indicates a high correspondence between the observed and estimated item metrics.

Subjective Beliefs and Effect Size Estimates

Figure 4 displays the average prior and final beliefs about the plausibility of the hypothesis of interest.[4] Researchers' prior beliefs about religiosity being positively related

[4] Note that prior beliefs about the plausibility of the effect were reported on a 7-pt scale instead of a 4-pt 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-pt scale into a 4-pt 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

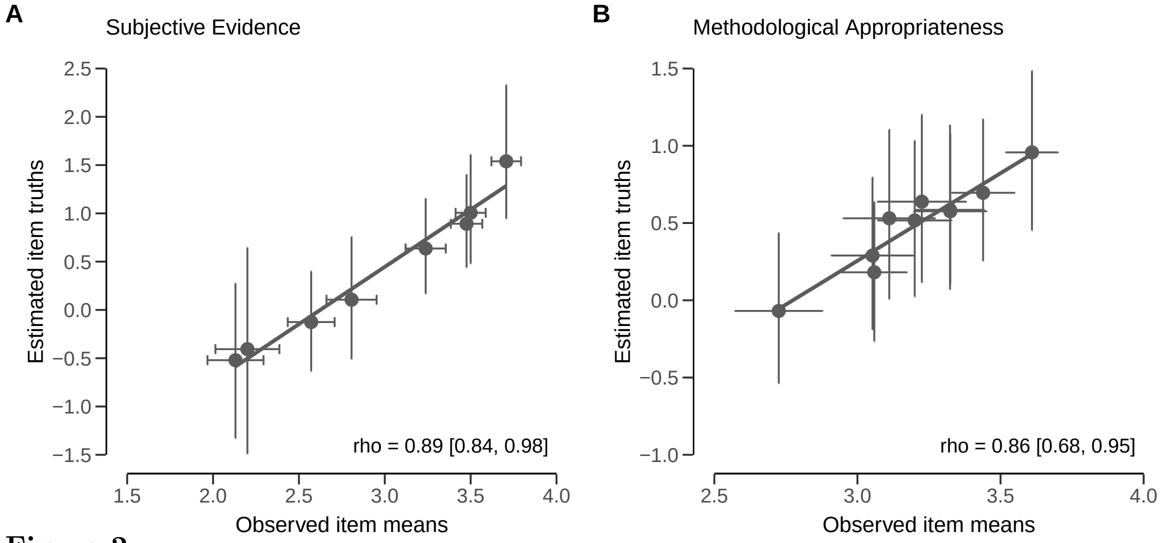


Figure 2

*Correlation between the estimated item truths and the observed item means for the subjective evidence subscale (panel **A**) and the methodological appropriateness subscale (panel **B**). Vertical error bars reflect the 95% credible interval of the estimated item truths and horizontal error bars reflect the standard error of the observed scores.*

Table 2

Descriptive Statistics for the Methodological Appropriateness Subscale for the Pilot Data

Label	M	SD	No answer
Sampling plan	3.06	0.74	19%
Statistical power	3.61	0.59	2.4%
Missing values	3.32	0.82	11.9%
Biased sample	3.20	0.83	16.7%
Study setting	3.32	0.77	19%
Reliability	3.05	0.93	9.5%
Validity	2.72	0.99	4.8%
Research design	3.22	1.00	4.8%
Missing variables	3.11	1.04	14.3%
Analysis	3.44	0.71	2.4%

Note. All items were answered on a 4-point scale.

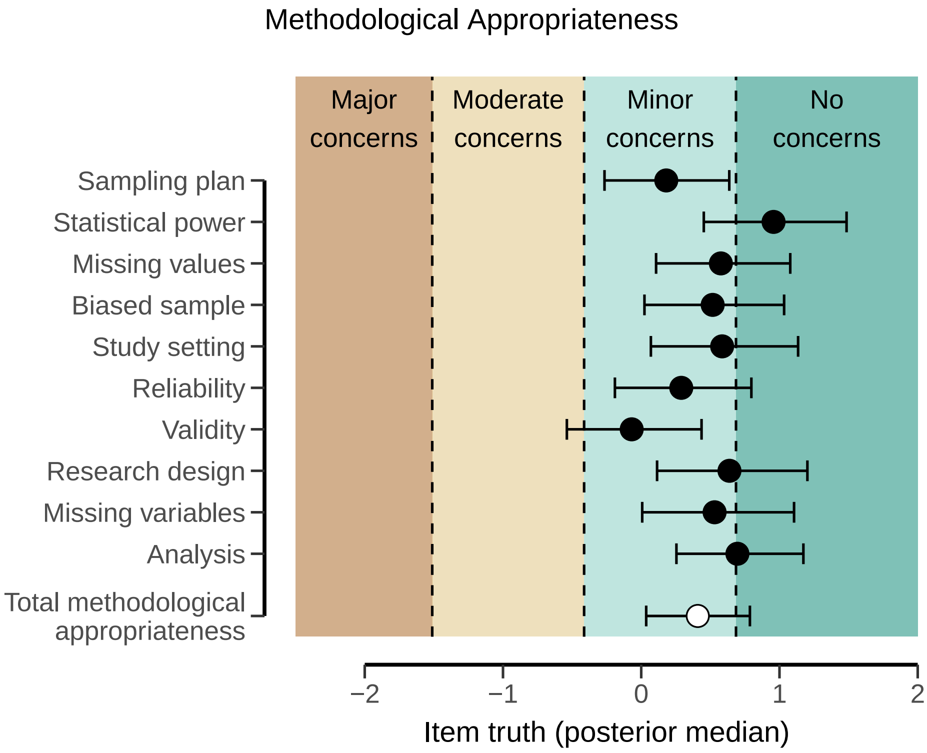


Figure 3

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

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 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 Silberzahn et al. (2018) and Hoogeveen, Sarafoglou, Aczel, et al. (2023), 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

randomly assigned to become 3 or 4.

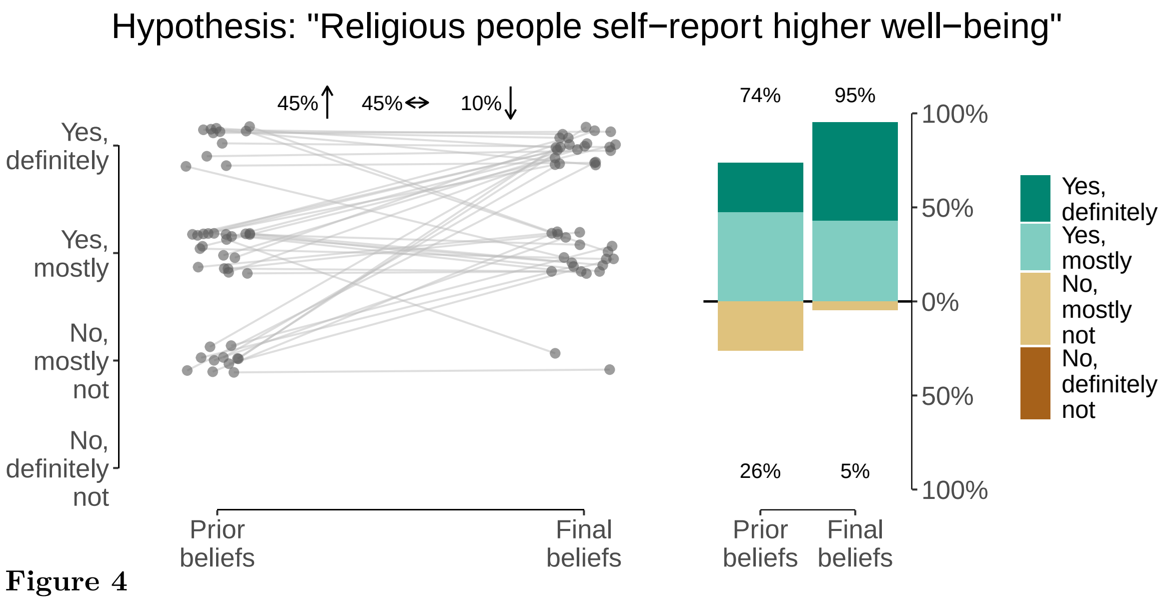


Figure 4

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 percent of the teams considered the hypothesis more likely after having analyzed the data than prior to seeing the data, 10% considered the hypothesis less likely having analyzed 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.

analyzing the data. In addition, we evaluated whether the effect sizes were related to the estimates of individual skepticism (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 skepticism, reflecting that analysis teams who found lower effect sizes were subjectively more skeptical (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 skepticism. 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

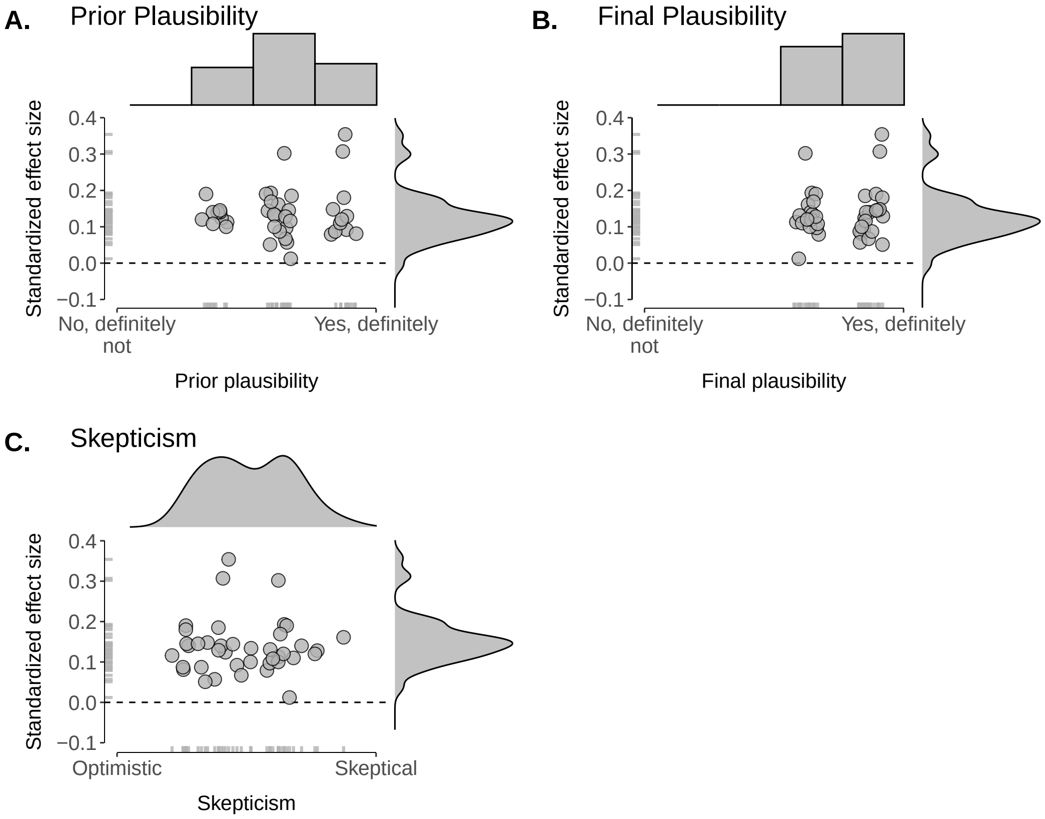


Figure 5

Reported effect sizes (beta coefficients) and subjective beliefs about the likelihood of the hypothesis. A. shows the relation between effect size and prior beliefs for the research question, B. shows the relation between effect size and final beliefs for the research question, and C. shows the relation between effect size and the analysis teams' level of skepticism regarding the evidence. In panels A and 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.

and estimated skepticism.

The correlations are visualized in Figure 5. We obtained moderate evidence *against* a positive relation between prior beliefs about the plausibility of the hypothesis and the reported effect sizes: $\text{BF}_{+0} = 0.13$; $\text{BF}_{0+} = 7.94$, $\rho_s =$ -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: $\text{BF}_{+0} = 0.08$; $\text{BF}_{0+} = 12.11$, $\rho_s =$ -0.27, 95% credible interval [-0.53, 0.06].[5] Finally, we found anecdotal evidence against a negative relation between estimated skepticism on the SEES and reported effect sizes: $\text{BF}_{+0} = 0.35$; $\text{BF}_{0+} = 2.85$, $\rho = 0.00$, 95% credible interval [-0.30, 0.28].

As mentioned in Hoogeveen, Sarafoglou, Aczel, et al. (2023), 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 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 skepticism. 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 skepticism.[6]

[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 $\text{BF}_{+0} = 0.08$; $\text{BF}_{0+} = 12.23$, $\rho_s =$ -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 $\text{BF}_{+0} = 1.09$; $\text{BF}_{0+} = 0.91$, $\rho_s = 0.19$, 95% credible interval [-0.03, 0.38].

[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 (Hoogeveen, Sarafoglou, Aczel, et al., 2023).

Moreover, the long time period between conducting the analyses and completing the SEES prevent 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.

Limitations

The applicability of the SEES may vary depending on the nature of the multi-analyst project. In a typical scenario, multi-analyst 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 multi-analyst projects essentially eliminate opportunity for analysts to choose which variables to assess. For instance, commentaries published on the Many-Analysts Religion Project (e.g., Edelsbrunner et al., 2023; Murphy & Martinez, 2023) argued that project leaders should have reduced ambiguity as much as possible to avoid 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* behaviors and/or beliefs benefit *specific* populations' well-being or health in *specific* contexts" (Murphy & Martinez, 2023, p.2). In a multi-analyst project which aims to reduce ambiguity 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 multi-analyst study is to capture (the consequences of) different principled decisions throughout the analytic process.

In addition, some data sets 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 Silberzahn et al. (2018), 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 data set, rather than on multiple data sets (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 multi-analyst 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 behavioral sciences than to researchers from other areas. If necessary, project leaders from multi-analyst 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, 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 item truths 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 favorable recovery for true item truths (i.e., wider posterior distributions) compared to the larger sample. The full results can be accessed in the online supplements at https://osf.io/4cesj.

Discussion

The present work introduces the Subjective Evidence Evaluation Survey (SEES) as a tool to systematically explore and quantify subjective measures of evidence in multi-analyst 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 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 multi-analyst project. Combined with the objective outcomes of the multi-analyst 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 multi-analyst 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.

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 multi-analyst 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 than 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. Silberzahn et al., 2018). We recommend using the outlined Bayesian cultural consensus theory model to analyze 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, especially when the number of participating analysis teams is low.

We contribute to the current literature about guidelines on multi-analyst studies (Aczel et al., 2021) by offering concrete advice on how to analyze and interpret (part of) the data obtained in multi-analyst projects. This, together with advancements on synthesizing objective outcome metrics across analyses based on the same data (e.g., Coretta et al., 2023; Kümpel & Hoffmann, 2022), can move the field beyond drawing conclusions based on (visual) inspection of the analysts' outcomes.

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 multi-analyst project. We hope to have inspired project leaders of multi-analyst projects to consider adding subjective evidence assessment to their future projects and thereby allowing for a more complete evaluation of the outcomes.

Disclosures

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 manuscript. 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.

Data and Materials

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/.

Ethical Approval

The expert consensus procedure was approved by the local ethics board of the University of Amsterdam (registration number: 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 multi-analyst studies can be analyzed 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 Euro voucher as compensation.

Author Contributions

Contributorship was documented with CRediT taxonomy using tenzing (Holcombe et al., 2020).

Conceptualization: A.S., S. Hoogeveen, and E.-J.W.

Data curation: A.S. and S. Hoogeveen

Formal analysis: A.S., S. Hoogeveen, and D.v.d.B.

Funding acquisition: A.S., S. Hoogeveen, and E.-J.W.

Survey creation: A.S., S. Hoogeveen, E.-J.W., B.A., C.J.A., T.A., R.B.-N., N.A.B., A.M.C., B.D., N.N.N.v.D., A.D., E.F., R.H., S. Hoffman, F.H., J.H., N.H.-K., J.I., M.J., M.K., E.L., J.-F.M., D.M., A.J.M., G.N., D.v.R., M.J.S., H.S.-K., D.R.S., D.J.S., B.A.S., A.H.S., B.S., D.T., F.T., E.L.U., W.V., and J.W.

Methodology: A.S., S. Hoogeveen, D.v.d.B., and E.-J.W.

Project administration: A.S., S. Hoogeveen, and E.-J.W.

Resources: A.S., S. Hoogeveen, and E.-J.W.

Software: A.S., S. Hoogeveen, and D.v.d.B.

Supervision: E.-J.W.

Validation: A.S., S. Hoogeveen, D.v.d.B., and E.-J.W.

Visualization: A.S. and S. Hoogeveen

Writing - original draft: A.S., S. Hoogeveen, D.v.d.B., and E.-J.W.

Writing - review & editing: A.S., S. Hoogeveen, D.v.d.B., E.-J.W., B.A., C.J.A., T.A., R.B.-N., N.A.B., A.M.C., B.D., N.N.N.v.D., A.D., E.F., R.H., S. Hoffman, F.H., J.H., N.H.-K., J.I., M.J., M.K., E.L., J.-F.M., D.M., A.J.M., G.N., D.v.R., M.J.S., H.S.-K., D.R.S., D.J.S., B.A.S., A.H.S., B.S., D.T., F.T., E.L.U., W.V., and J.W.

Conflicts of Interest

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of 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 program "META-REP: A Meta-scientific Programme to Analyse and Optimise 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 a 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).

References

Aczel, B., Szaszi, B., Nilsonne, G., van den Akker, O. R., Albers, C. J., van Assen, M. A., Bastiaansen, J. A., Benjamin, D., Boehm, U., Botvinik-Nezer, R., Bringmann, L. F., Busch, N. A., Caruyer, E., Cataldo, A. M., Cowan, N., Delios, A., van Dongen,

N. N., Donkin, C., van Doorn, J. B., … Wagenmakers, E.-J. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife*, *10*, e72185. https://doi.org/10.7554/eLife.72185

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharskỳ, Š., Benjamin, D., Chambers, C. D., Fisher, A., Gelman, A., Gernsbacher, M. A., et al. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, *4*(1), 4–6.

Anders, R., & Batchelder, W. H. (2012). Cultural Consensus Theory for multiple consensus truths. *Journal of Mathematical Psychology*, *56*, 452–469. https://doi.org/10.1016/j.jmp.2013.01.004

Anders, R., & Batchelder, W. H. (2015). Cultural Consensus Theory for the ordinal data case. *Psychometrika*, *80*(1), 151–181. https://doi.org/10.1007/s11336-013-9382-9

Atkinson, Q. D., Claessens, S., Fischer, K., Forsyth, G. L., Kyritsis, T., Wiebels, K., & Moreau, D. (2023). Being specific about generalisability. *Religion, Brain & Behavior*, *13*(3), 284–286. https://doi.org/10.1080/2153599X.2022.2070251

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., Chow, S.-M., de Jonge, P., Emerencia, A. C., Epskamp, S., Fisher, A. J., Hamaker, E. L., Kuppens, P., Lutz, W., Meyer, M. J., Moulder, R., Oravecz, Z., Riese, H., Rubel, J., … Bringmann, L. F. (2020). Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*, *137*, 110211. https://doi.org/10.1016/j.jpsychores.2020.110211

Batchelder, W. H., & Anders, R. (2012). Cultural Consensus Theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, *56*, 316–332. https://doi.org/10.1016/j.jmp.2012.06.002

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., Krypotos, A.-M., Lerche, V., Logan, G. D., Palmeri, T. J., van Ravenzwaaij, D., Servant, M., Singmann, H., Starns, J. J., Voss, A., Wiecki, T. V., Matzke, D., & Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the Diffusion Decision

Model: Expert advice and recommendations. *Journal of Mathematical Psychology*, *87*, 46–75. https://doi.org/10.1016/j.jmp.2018.09.004

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., . . . Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H. V., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., . . . Żółtak, T. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, *119*(44), e2203150119. https://doi.org/10.1073/pnas.2203150119

Briner, R. B., Denyer, D., et al. (2012). Systematic review and evidence synthesis as a practice and scholarship tool. *Handbook of evidence-based management: Companies, classrooms and research*, 112–129.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*. https://doi.org/10.18637/jss.v076.i01

Colvin, C. J., Garside, R., Wainwright, M., Munthe-Kaas, H., Glenton, C., Bohren, M. A., Carlsen, B., Tunçalp, Ö., Noyes, J., Booth, A., Rashidian, A., Flottorp, S., & Lewin, S. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings—paper 4: How to assess coherence. *Implementation Science*, *13*(1), 13. https://doi.org/10.1186/s13012-017-0691-8

Coretta, S., Casillas, J. V., Roessig, S., Franke, M., Ahn, B., Al-Hoorie, A. H., Al-Tamimi, J., Alotaibi, N. E., AlShakhori, M. K., Altmiller, R. M., Arantes, P., Athanasopoulou, A., Baese-Berk, M. M., Bailey, G., Sangma, C. B. A., Beier, E. J., Benavides, G. M., Benker, N., BensonMeyer, E. P., . . . Roettger, T. B. (2023). Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231162567. https://doi.org/10.1177/25152459231162567

Critical Appraisal Skills Programme. (2018a). *CASP Cohort Study Checklist* (tech. rep.).

Critical Appraisal Skills Programme. (2018b). *CASP Qualitative Checklist* (tech. rep.).

de Vrieze, J. (2018). The metawars. *Science*, *361*(6408), 1184–1188. https://doi.org/10.1126/science.361.6408.1184

Donnelly, S., Brooks, P. J., & Homer, B. D. (2019). Is there a bilingual advantage on interference-control tasks? A multiverse meta-analysis of global reaction time and interference cost. *Psychonomic Bulletin & Review*, *26*(4), 1122–1147. https://doi.org/10.3758/s13423-019-01567-z

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Krypotos, A.-M., Kupitz, C. N., Leite, F. P., Lerche, V., Lin, Y.-S., Logan, G. D., Palmeri, T. J., Starns, J. J., Trueblood, J. S., van Maanen, L., . . . Donkin, C. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review*, *26*(4), 1051–1069. https://doi.org/10.3758/s13423-017-1417-2

Edelsbrunner, P. A., Sebben, S., Frisch, L. K., Schüttengruber, V., Protzko, J., & Thurn, C. M. (2023). How to understand a research question – A challenging first step in setting up a statistical model. *Religion, Brain & Behavior*, *13*(3), 306–309. https://doi.org/10.1080/2153599X.2022.2070258

Fillard, P., Descoteaux, M., Goh, A., Gouttard, S., Jeurissen, B., Malcolm, J., Ramirez-Manzanares, A., Reisert, M., Sakaie, K., Tensaouti, F., Yo, T., Mangin, J.-F., & Poupon, C. (2011). Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. *NeuroImage*, *56*(1), 220–234. https://doi.org/10.1016/j.neuroimage.2011.01.032

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. *American Scientist*, *102*(6), 460.

Gould, E., Fraser, H. S., Parker, T. H., Nakagawa, S., Griffith, S. C., Vesk, P. A., Fidler, F., Hamilton, D. G., Abbey-Lee, R. N., Abbott, J. K., Aguirre, L. A., Alcaraz, C., Aloni, I., Altschul, D., Arekar, K., Atkins, J. W., Atkinson, J., Baker, C., Barrett, M., . . . Zitomer, R. A. (2023). Same data, different analysts: Variation in effect sizes due to analytical decisions in ecology and evolutionary biology. https://doi.org/10.32942/X2GG62

Hanel, P. H. P., & Zarzeczna, N. (2023). From multiverse analysis to multiverse operationalisations: 262,143 ways of measuring well-being. *Religion, Brain & Behavior*, *13*(3), 309–313. https://doi.org/10.1080/2153599X.2022.2070259

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623.

Holcombe, A. O., Kovacs, M., Aust, F., & Aczel, B. (2020). Documenting contributions to scholarly articles using CRediT and tenzing. *PLoS One*, *15*(12), e0244611.

Holzmeister, F., Johannesson, M., Böhm, R., Almenberg, A. D., Huber, J., & Kirchler, M. (2024). Heterogeneity in effect size estimates: Empirical evidence and practical implications. https://doi.org/10.31222/osf.io/583un

Hoogeveen, S., Berkhout, S., Gronau, Q. F., Wagenmakers, E.-J., & Haaf, J. M. (2023). Improving statistical analysis in team science: The case of a Bayesian multiverse of Many Labs 4. https://doi.org/10.31234/osf.io/cb9er

Hoogeveen, S., Sarafoglou, A., Aczel, B., Aditya, Y., Alayan, A., Allen, P., Altay, S., Alzahawi, S., Amir, Y., Anthony, F.-V., Appiah, O., Atkinson, Q. D., Baimel, A., Balkaya-Ince, M., Balsamo, M., Banker, S., Bartoš, F., Becerra, M., Beffara, B., . . . Wagenmakers, E.-J. (2023). A many-analysts approach to the relation between religiosity and well-being. *Religion, Brain & Behavior*, *13*(3), 237–283. https://doi.org/10.1080/2153599X.2022.2070255

Hoogeveen, S., Sarafoglou, A., van Elk, M., & Wagenmakers, E.-J. (2023). Many-Analysts Religion Project: Reflection and conclusion. *Religion, Brain & Behavior*, *13*(3), 356–363. https://doi.org/10.1080/2153599X.2022.2070263

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M., & Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, *59*(3), 944–960. https://doi.org/10.1111/ecin.12992

Klau, S., Hoffmann, S., Patel, C. J., Ioannidis, J. P., & Boulesteix, A.-L. (2021). Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework. *International Journal of Epidemiology*, *50*(1), 266–278. https://doi.org/10.1093/ije/dyaa164

Krypotos, A.-M., Klein, R., & Jong, J. (2023). Resolving religious debates through a multiverse approach. *Religion, Brain & Behavior*, *13*(3), 318–320. https://doi.org/10.1080/2153599X.2022.2070261

Kucharský, Š., Tran, N.-H., Veldkamp, K., Raijmakers, M., & Visser, I. (2021). Hidden Markov Models of Evidence Accumulation in Speeded Decision Tasks. *Computational Brain & Behavior*, *4*(4), 416–441. https://doi.org/10.1007/s42113-021-00115-0

Kümpel, H., & Hoffmann, S. (2022). A formal framework for generalized reporting methods in parametric settings. https://doi.org/10.48550/arXiv.2211.02621

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Lewin, S., Bohren, M., Rashidian, A., Munthe-Kaas, H., Glenton, C., Colvin, C. J., Garside, R., Noyes, J., Booth, A., Tunçalp, Ö., Wainwright, M., Flottorp, S., Tucker, J. D., & Carlsen, B. (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. *Implementation Science*, *13*(1), 10. https://doi.org/10.1186/s13012-017-0689-2

Maier-Hein, K. H., Neher, P. F., Houde, J.-C., Côté, M.-A., Garyfallidis, E., Zhong, J., Chamberland, M., Yeh, F.-C., Lin, Y.-C., Ji, Q., Reddick, W. E., Glass, J. O., Chen, D. Q., Feng, Y., Gao, C., Wu, Y., Ma, J., He, R., Li, Q., . . . Descoteaux, M. (2017). The challenge of mapping the human connectome based on diffusion tractography. *Nature Communications*, *8*(1), 1349. https://doi.org/10.1038/s41467-017-01285-x

Mathur, M. B., Covington, C., & VanderWeele, T. J. (2023). Variation across analysts in statistical significance, yet consistently small effect sizes. *Proceedings of the National Academy of Sciences*, *120*(3), e2218957120. https://doi.org/10.1073/pnas.2218957120

McKenna, H. P. (1994). The Delphi technique: A worthwhile research approach for nursing? *Journal of Advanced Nursing*, *19*(6), 1221–1225. https://doi.org/10.1111/j.1365-2648.1994.tb01207.x

McNamara, A. A. (2023). The impact (or lack thereof) of analysis choice on conclusions with Likert data from the Many Analysts Religion Project. *Religion, Brain & Behavior*, *13*(3), 324–326. https://doi.org/10.1080/2153599X.2022.2070256

Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Neusüss, S., Razen, M., & Weitzel, U. (2021). Non-standard errors. https://doi.org/10.2139/ssrn.3961574

Modecki, K. L., Low-Choy, S., Uink, B. N., Vernon, L., Correia, H., & Andrews, K. (2020). Tuning into the real effect of smartphone use on parenting: A multiverse analysis. *Journal of Child Psychology and Psychiatry*, *61*(8), 855–865. https://doi.org/10.1111/jcpp.13282

Murphy, J., & Martinez, N. (2023). Quantifying religiosity: A comparison of approaches based on categorical self-identification and multidimensional measures of religious activity. *Religion, Brain & Behavior*, *13*(3), 327–329. https://doi.org/10.1080/2153599X.2022.2070252

Odenbaugh, J., & Alexandrova, A. (2011). Buyer beware: Robustness analyses in economics and biology. *Biology & Philosophy*, *26*(5), 757–771. https://doi.org/10.1007/s10539-011-9278-y

Oravecz, Z., Anders, R., & Batchelder, W. H. (2015). Hierarchical Bayesian Modeling for Test Theory Without an Answer Key. *Psychometrika*, *80*(2), 341–364. https://doi.org/10.1007/s11336-013-9379-4

Oravecz, Z., Vandekerckhove, J., & Batchelder, W. H. (2014). Bayesian Cultural Consensus Theory. *Field Methods*, *26*(3), 207–222. https://doi.org/10.1177/1525822X13520280

Oza, A. (2023). Reproducibility trial: 246 biologists get different results from same data sets. *Nature*, *622*(7984), 677–678. https://doi.org/10.1038/d41586-023-03177-1 Bandiera_abtest: a Cg_type: News Subject_term: Ecology, Peer review, Evolution, Scientific community, Research data

Palpacuer, C., Hammas, K., Duprez, R., Laviolle, B., Ioannidis, J. P. A., & 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 Medicine*, *17*(1), 174. https://doi.org/10.1186/s12916-019-1409-3

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, *68*(9), 1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

Pearson, H. I., Lo, R. F., & Sasaki, J. Y. (2023). How do culture and religion interact worldwide? A cultural match approach to understanding religiosity and well-being in the Many Analysts Religion Project Hannah I. Pearson1, *Ronda F. Lo2, Joni Y. Sasaki. *Religion, Brain & Behavior*, *13*(3), 329–336. https://doi.org/10.1080/2153599X.2022.2070265

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*(2), 313–338. https://doi.org/10.1525/aa.1986.88.2.02a00020

Ross, R. M., Sulik, J., Buczny, J., & Schivinski, B. (2023). Many analysts and few incentives. *Religion, Brain & Behavior*, *13*(3), 336–339. https://doi.org/10.1080/2153599X.2022.2070248

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., . . . McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398–8403. https://doi.org/10.1073/pnas.1915006117

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, *26*(1), 103–126. https://doi.org/10.1037/met0000275

Schreiner, M. R., Mercier, B., Frick, S., Wiwad, D., Schmitt, M. C., Kelly, J. M., & Quevedo Pütter, J. (2023). Measurement issues in the Many Analysts Religion Project. *Religion, Brain & Behavior*, *13*(3), 339–341. https://doi.org/10.1080/2153599X.2022.2070260

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C. M., van Assen, M. A. L. M., Liu, Y., Althoff, T., Heer, J., Kale, A., Mohamed, Z., Amireh, H., Venkatesh Prasad, V., Bernstein, A., Robinson, E., Snellman, K., Amy Sommer, S., Otner, S. M. G., Robinson, D., . . . Luis Uhlmann, E. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes*, *165*, 228–249. https://doi.org/10.1016/j.obhdp.2021.02.003

Scientific Pandemic Influenza Group on Modelling. (2020). SPI-M-O: Consensus statement on COVID-19, 8 October 2020.

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

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahnik, Š., Bai, F., Bannard, C., Bonnier, E., et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Smith, E. (2023). Individual-level versus country-level moderation. *Religion, Brain & Behavior*, *13*(3), 342–344. https://doi.org/10.1080/2153599X.2022.2070246

Spencer, L., Ritchie, J., Lewis, J., Dillon, L., et al. (2004). *Quality in qualitative evaluation: A framework for assessing research evidence* (tech. rep.). Government Chief Social Researcher's Office.

Starns, J. J., Cataldo, A. M., Rotello, C. M., Annis, J., Aschenbrenner, A., Bröder, A., Cox, G., Criss, A., Curl, R. A., Dobbins, I. G., Dunn, J., Enam, T., Evans, N. J., Farrell, S., Fraundorf, S. H., Gronlund, S. D., Heathcote, A., Heck, D. W., Hicks, J. L., ... Wilson, J. (2019). Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. *Advances in Methods and Practices in Psychological Science*, *2*(4), 335–349. https://doi.org/10.1177/2515245919869583

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2018). Validating Bayesian Inference Algorithms with Simulation-Based Calibration. https://doi.org/10.48550/arXiv.1804.06788

Trübutschek, D., Yang, Y.-F., Gianelli, C., Cesnaite, E., Fischer, N. L., Vinding, M. C., Marshall, T. R., Algermissen, J., Pascarella, A., Puoliväli, T., Vitale, A., Busch,

N. A., & Nilsonne, G. (2023). EEGManyPipelines: A large-scale, grassroots multi-analyst study of electroencephalography analysis practices in the wild. *Journal of Cognitive Neuroscience*, 1–8. https://doi.org/10.1162/jocn_a_02087

van den Bergh, D., Bogaerts, S., Spreen, M., Flohr, R., Vandekerckhove, J., Batchelder, W. H., & Wagenmakers, E.-J. (2020). Cultural Consensus Theory for the evaluation of patients' mental health scores in forensic psychiatric hospitals. *Journal of Mathematical Psychology*, *98*, 102383. https://doi.org/10.1016/j.jmp.2020.102383

van Elk, M., Matzke, D., Gronau, Q., Guang, M., Vandekerckhove, J., & Wagenmakers, E.-J. (2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. *Frontiers in Psychology*, *6*, 1365. https://doi.org/10.3389/fpsyg.2015.01365

van Lissa, C. J. (2023). Complementing preregistered confirmatory analyses with rigorous, reproducible exploration using machine learning. *Religion, Brain & Behavior*, *13*(3), 347–351. https://doi.org/10.1080/2153599X.2022.2070254

van Dongen, N. N. N., van Doorn, J. B., Gronau, Q. F., van Ravenzwaaij, D., Hoekstra, R., Haucke, M. N., Lakens, D., Hennig, C., Morey, R. D., Homer, S., Gelman, A., Sprenger, J., & Wagenmakers, E.-J. (2019). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician*, *73*(sup1), 328–339. https://doi.org/10.1080/00031305.2019.1565553

Veronese, M., Rizzo, G., Belzunce, M., Schubert, J., Searle, G., Whittington, A., Mansur, A., Dunn, J., Reader, A., & Gunn, R. N. (2021). Reproducibility of findings in modern PET neuroimaging: Insight from the NRM2018 grand challenge. *Journal of Cerebral Blood Flow & Metabolism*, *41*(10), 2778–2796. https://doi.org/10.1177/0271678X211015101

Vogel, V., Prenoveau, J., Kelchtermans, S., Magyar-Russell, G., McMahon, C., Ingendahl, M., & Schaumans, C. B. C. (2023). Different facets, different results: The importance of considering the multidimensionality of constructs. *Religion, Brain & Behavior*, *13*(3), 351–356. https://doi.org/10.1080/2153599X.2022.2070262

Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, *605*(7910), 423–425. https://doi.org/10.1038/d41586-022-01332-8

Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2023). Facing the unknown unknowns of data analysis. *Current Directions in Psychological Science*, 09637214231168565. https://doi.org/10.1177/09637214231168565

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.31219/osf.io/umq8d

Appendix A

Subjective Evidence Evaluation Survey: Instructions To Project Leaders

The Subjective Evidence Evaluation Survey (SEES) has been developed to facilitate and extend the interpretation of evidence for a given research hypothesis in multi-analyst projects. In a multi-analyst project, multiple research teams are invited to independently analyze the same data and address the key research question (Silberzahn & Uhlmann, 2015; Silberzahn et al., 2018). The core idea of a multi-analyst 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.

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.

Post-Analysis Phase

Subjective Evidence Subscale

The subjective evidence subscale consists of 8 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 A1.

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."

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 skeptical to favorable.

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.*

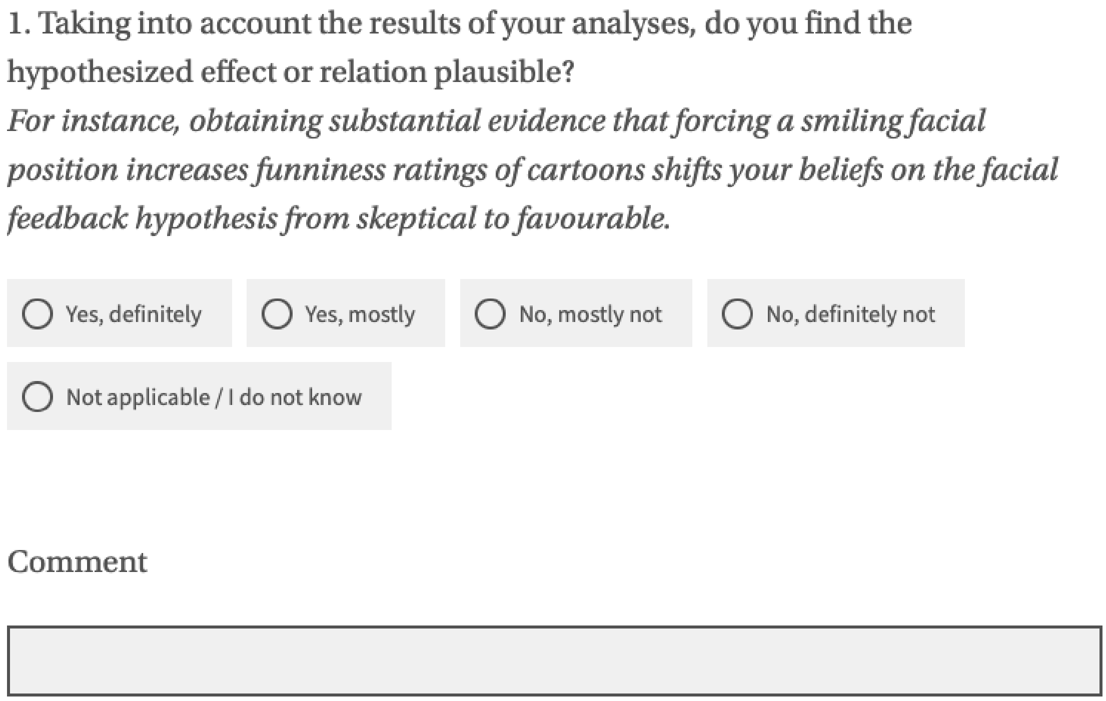
4. Does your analysis based on the observed data provide substantial evidence *against* the hypothesized effect or relation? *For instance, evidence points in the opposite direction than hypothesized, or the evidence favors the null hypothesis.*

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 vs. 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.*



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 skeptical to favourable.

○ Yes, definitely ○ Yes, mostly ○ No, mostly not ○ No, definitely not

○ Not applicable / I do not know

Comment

Figure A1

Example of the presentation of the subjective evidence subscale, item 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 A2.

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."

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.*

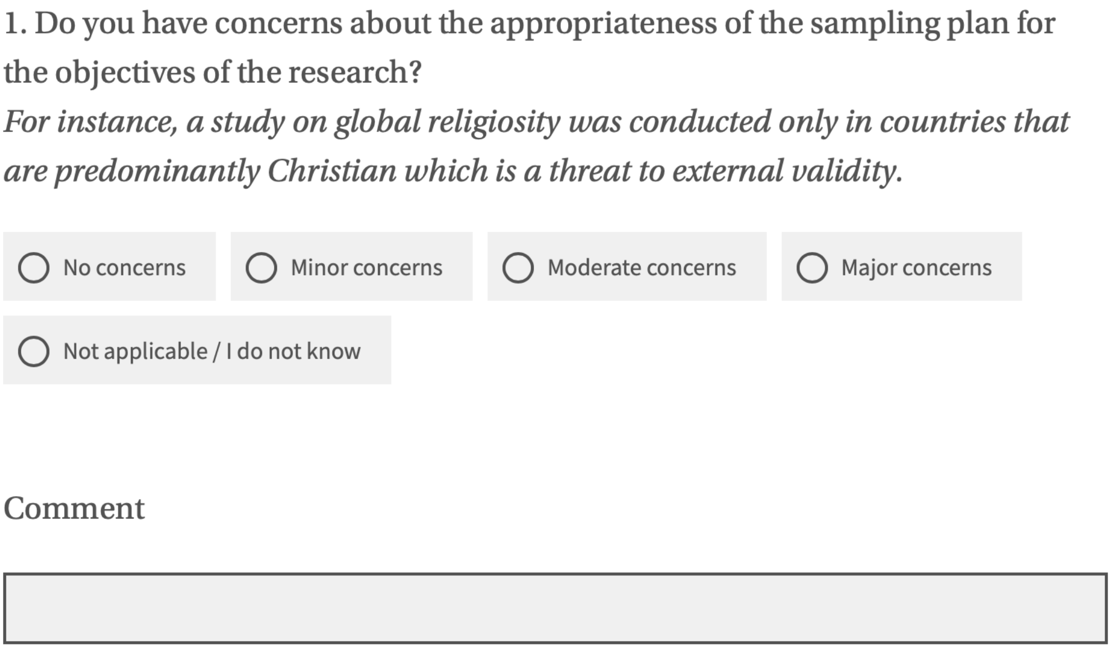
3. Do you have concerns about missing values on the relevant variables? *For instance, there are too many missing values to draw a statistically valid conclusion, or the pattern of missing values appears non-random.*

4. Do particular sample characteristics (e.g., age, gender, 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.*

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.*



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.

○ No concerns ○ Minor concerns ○ Moderate concerns ○ Major concerns
○ Not applicable / I do not know

Comment

Figure A2

Example of the presentation of the methodological concerns subscale, item 1.

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.*

Background Information

Survey Development

The SEES was developed in collaboration with 37 experts in relevant scientific areas following a preregistered 'reactive-Delphi' expert consensus procedure (McKenna, 1994) as implemented in Aczel et al. (2021) and Aczel et al. (2020). Selected areas of expertise included multi-analyst 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. Please note that we preregistered that only items with a median recommendation ration 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. All items received approval from panel members during the discussion rounds.

Adaptations

We encourage project leaders from multi-analyst studies to use the SEES flexibly and reword the examples for the items to the specific field of study if necessary. In addition, some data sets 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 data set, rather than on multiple data sets (e.g., stemming from multiple experiments), project leaders could present the methodological appropriateness subscale only once to the teams.

Proposed Analysis

The SEES data can be analyzed in a cultural consensus theory model (Anders & Batchelder, 2012; Anders & Batchelder, 2015; Batchelder & Anders, 2012; Romney et al., 1986; van den Bergh et al., 2020), which allows one to synthesize responses and capture the collective opinion about the subjective evidence in multi-analyst projects. A comprehensive description of the cultural consensus theory model applied to the SEES can be found appendix B of the manuscript.

Appendix B

Cultural Consensus Theory Model

Here we describe the adapted version of the latent truth rater model (Anders & Batchelder, 2015; van den Bergh et al., 2020) which we use to synthesize responses and capture collective opinions on the two subscales of the SEES.

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

[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.

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 item truth 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 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 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 B1 and explained mathematically below:

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

Latent Item Truth

The appraisal of an item $y_{r,i}$ comprises two latent components: the latent item truth (θ_i) and the item difficulty (κ_i). Within the context of a multi-analyst project, the primary focus is on estimating the latent item truth as it represents the items' true location on an assumed underlying unidimensional scale. The assumption is that, depending on the

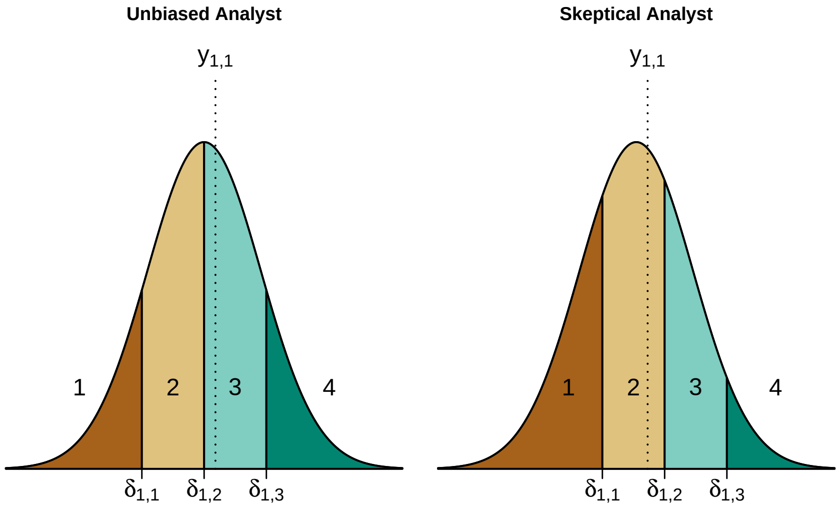


Figure B1

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 the left panel) would respond 'yes, mostly', while the skeptical analyst ($\beta > 0.5$; shown in the right panel) would respond 'no, mostly not'.

item difficulty, the latent item truth 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 item truth 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.

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 skepticism). 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, $\lambda_c = \text{logit}\left(\frac{c}{C}\right)$. In the scenario of an entirely unbiased analyst, their shift parameter would be $\beta_r = 0$, suggesting a neutral inclination toward selecting values on the response scale (see the left panel in Figure B1). 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 skeptical responses. To illustrate this shift, see the right panel in Figure B1 which depicts the latent probability distribution across predicted Likert scores for a skeptical 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 modeled 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:

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

Prior Distributions

We based our prior distributions on the suggestions provided by van den Bergh et al. (2020) 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 B2.

The applicability probability π_i for each item is assumed to be drawn from a beta distribution that mildly favors items being considered appropriate. The remaining parameters are assumed to be drawn from normal distributions. Due to identifiability constraints discussed in van den Bergh et al. (2020), the group-level mean for the item difficulty κ is fixed to 1:

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

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

$$\begin{aligned}\mu_\theta &\sim \text{Normal}(0, 0.25^2)\\ \mu_\beta &\sim \text{Normal}(0, 0.25^2).\end{aligned}$$

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

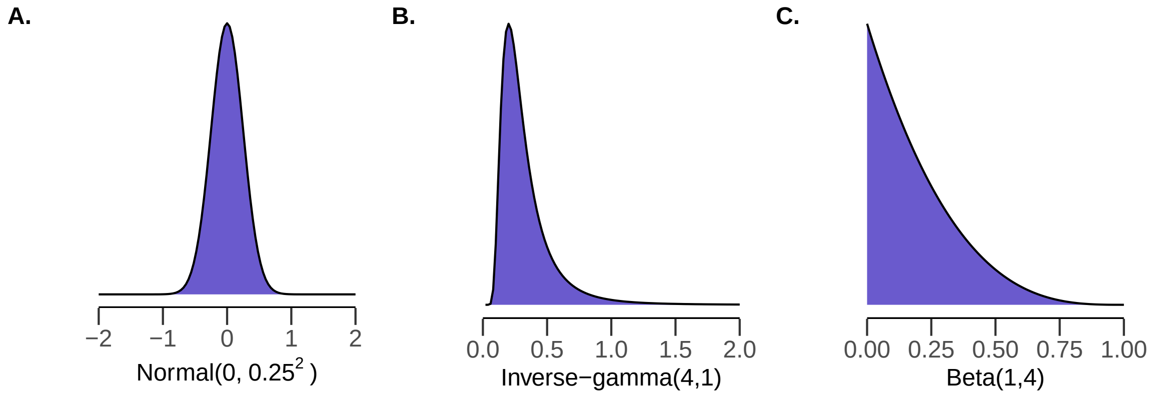


Figure B2

*Visualization of the group-level mean prior distributions (panel **A**), the group-level standard deviation prior distributions (panel **B**), and the prior distribution on the applicability probabilities (panel **C**) used in the SEES model.*

$$\sigma_\theta \sim \text{Inverse-Gamma}(4, 1)$$

$$\sigma_\kappa \sim \text{Inverse-Gamma}(4, 1)$$

$$\sigma_\beta \sim \text{Inverse-Gamma}(4, 1).$$

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 item truth. 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 by van den Bergh et al. (2020), 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.

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 B3 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 behavioral data and constitutes another step in the model validation process. Predictions from an adequate model should resemble the behavioral data. In Figure B4, 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 by Kucharský et al. (2021), based on the recommendations by Talts et al. (2018) and Schad et al. (2021). We conducted these checks for samples of size $n = 42$ and $n = 20$ with satisfactory model performance. The full results can be accessed in the online supplements.

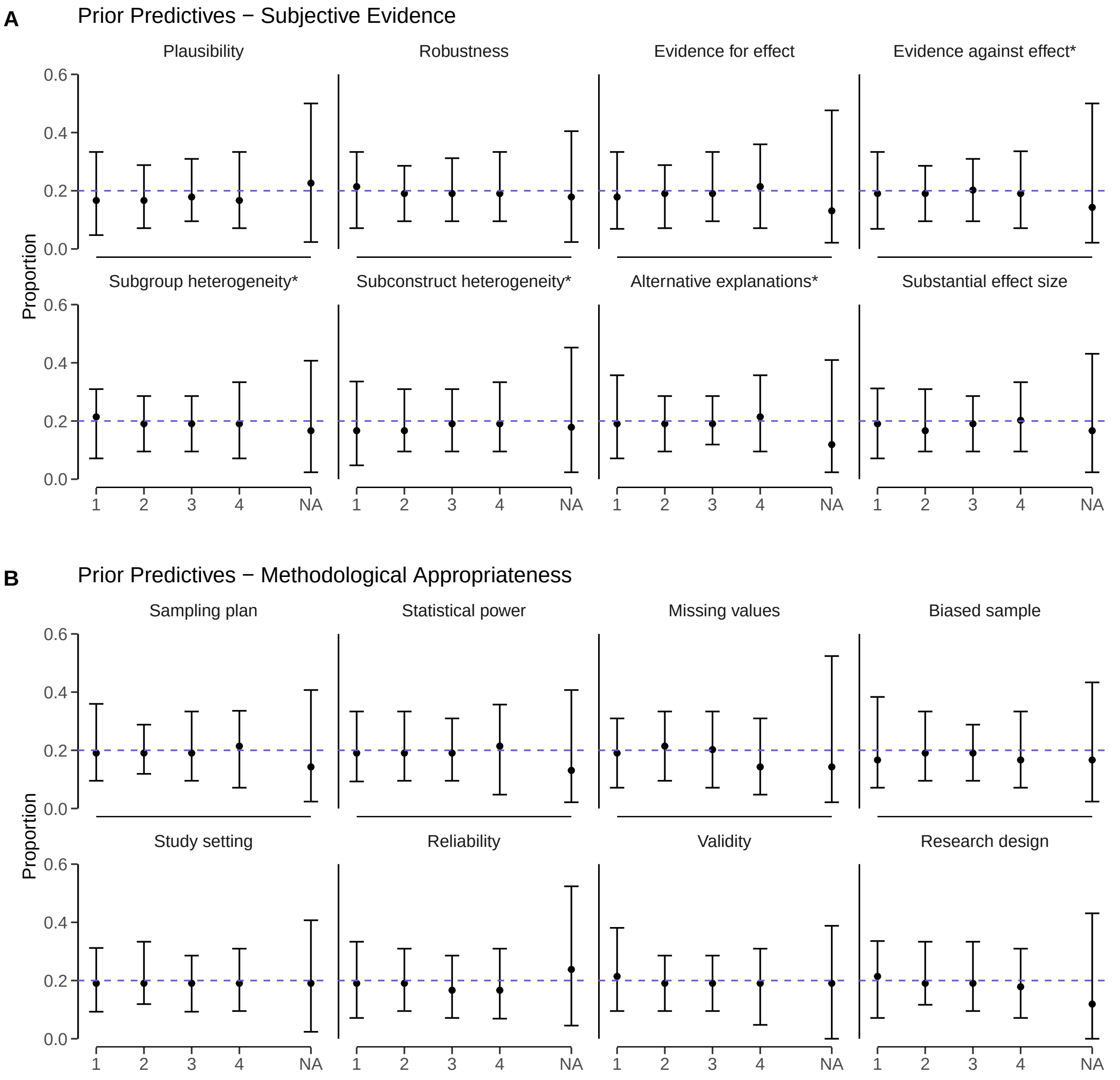


Figure B3

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.

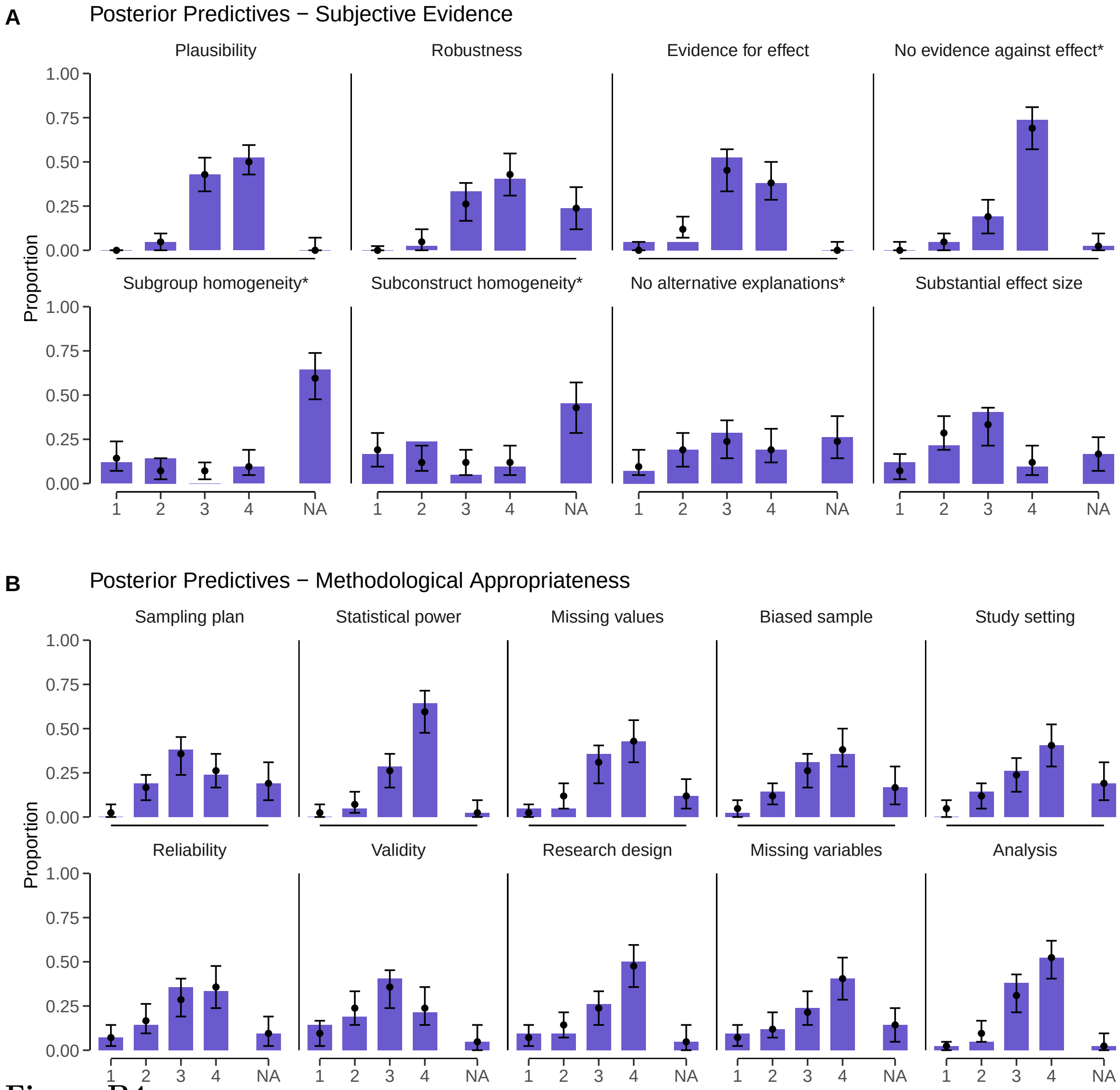


Figure B4

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.