nature human behaviour

PERSPECTIVE

https://doi.org/10.1038/s41562-021-01211-8

Seven steps toward more transparency in statistical practice

Eric-Jan Wagenmakers[1]✉, Alexandra Sarafoglou[1], Sil Aarts[2], Casper Albers[3], Johannes Algermissen[4], Štěpán Bahník[5], Noah van Dongen[1], Rink Hoekstra[6], David Moreau[7], Don van Ravenzwaaij[8], Aljaž Sluga[9], Franziska Stanke[10], Jorge Tendeiro[8,11] and Balazs Aczel[12]

We argue that statistical practice in the social and behavioural sciences benefits from transparency, a fair acknowledgement of uncertainty and openness to alternative interpretations. Here, to promote such a practice, we recommend seven concrete statistical procedures: (1) visualizing data; (2) quantifying inferential uncertainty; (3) assessing data preprocessing choices; (4) reporting multiple models; (5) involving multiple analysts; (6) interpreting results modestly; and (7) sharing data and code. We discuss their benefits and limitations, and provide guidelines for adoption. Each of the seven procedures finds inspiration in Merton's ethos of science as reflected in the norms of communalism, universalism, disinterestedness and organized scepticism. We believe that these ethical considerations—as well as their statistical consequences—establish common ground among data analysts, despite continuing disagreements about the foundations of statistical inference.

A superficial assessment of the published literature suggests that statisticians rarely agree on anything. Different schools—mostly frequentists, likelihoodists and Bayesians—have fought one another tooth and nail for decades, debating the meaning of 'probability', arguing about the role of prior knowledge, disputing the value of objective versus subjective analyses and disagreeing about the primary goal of inference itself: whether researchers should control error rates, update beliefs, or make coherent decisions. Fundamental disagreement exists not only between the different statistical schools, but also within the same school. For example, within the frequentist school, there is the perennial debate between those who seek to test hypotheses through *P* values and those who emphasize estimation through confidence intervals; and within the Bayesian school, Jack Good's claim that there are 46,656 varieties of Bayesians may prove to be an underestimate[1] (but see also ref. [2]).

The disagreement also manifests itself in practical application whenever multiple statisticians and practitioners of statistics find themselves independently analysing the same dataset. Specifically, recent 'multiple-analyst' articles show that statisticians rarely use the same analysis, and they often draw different conclusions, even for the exact same dataset and research question[3–7]. Deep disagreement is also exhibited by contradictory guidelines on *P* values[8–13]. Should practitioners avoid the phrase 'statistically significant'? Should they lower the *P* value thresholds, or justify them, or abandon *P* values altogether? And, if *P* values are abandoned, what should replace them? With statisticians fighting over these fundamental issues, users of applied statistics may be forgiven for adopting a wait-and-see attitude and carrying on as usual.

In this Perspective, we claim that, besides the numerous disputes and outstanding arguments, statisticians might agree on a set of scientific norms. We bring these norms to the fore, as we believe that they have considerable relevance for the practice of statistics in the social and behavioural sciences. The norms that we believe should guide statistical practice are communalism, universalism, disinterestedness and organized scepticism, which are the four scientific norms proposed by Merton[14] (originally published in 1942; see Box 1 for a detailed overview of the Mertonian norms and Box 2 for an overview of how each statistical procedure discussed here fulfills these norms).

Visualizing data

Description. By visualizing data, researchers can graphically represent key aspects of the observed data as well as important properties of the statistical model applied.

Benefits and examples. Data visualization is important in all phases of the statistical workflow. In exploratory data analysis, data visualization helps researchers to formulate new theories and hypotheses[15]. In model assessment, data visualization supports the detection of model misfit and guides the development of appropriate statistical models[16–20]. Finally, once the analysis is complete, visualization of data and model fit is arguably the most effective way to communicate the main findings to a scientific audience[21].

For an example of how data visualization facilitated the development of a new hypothesis, consider the famous map of the distribution of deaths from cholera created by London anaesthetist John Snow during the cholera outbreak in Soho, London, in September

[1]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands. [2]School for Public Health and Primary Care, Maastricht University, Maastricht, The Netherlands. [3]Heymans Institute of Psychological Research, University of Groningen, Groningen, The Netherlands. [4]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands. [5]Faculty of Business Administration, Prague University of Economics, Prague, Czech Republic. [6]Department of Educational Science, University of Groningen, Groningen, The Netherlands. [7]School of Psychology and Centre for Brain Research, The University of Auckland, Auckland, New Zealand. [8]Department of Psychology, University of Groningen, Groningen, The Netherlands. [9]Rotterdam School of Management, Erasmus University Rotterdam, Rotterdam, The Netherlands. [10]Department of Psychology, University of Münster, Münster, Germany. [11]Office of Research and Academia-Government-Community Collaboration Education and Research Center for Artificial Intelligence and Data Innovation, Hiroshima University, Hiroshima, Japan. [12]Institute of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary.
✉e-mail: ej.wagenmakers@gmail.com

Box 1 | Merton's ethos of science

Merton[14] proposed that scientific ethos is characterized by the following four norms:

(1) Communalism. "The substantive findings of science are a product of social collaboration and are assigned to the community. […] Property rights in science are whittled down to a bare minimum by the rationale of the scientific ethic. […] The institutional conception of science as part of the public domain is linked with the imperative for communication of findings. Secrecy is the antithesis of this norm; full and open communication its enactment." (pages 273–274 of ref. [14]).

(2) Universalism. "[T]ruth-claims, whatever their source, are to be subjected to preestablished impersonal criteria: consonant with observation and with previously confirmed knowledge. The acceptance or rejection of claims entering the lists of science is not to depend on the personal or social attributes of their protagonist; his race, nationality, religion, class, and personal qualities are as such irrelevant." (page 270 of ref. [14]).

(3) Disinterestedness. "Science, as is the case with professions in general, includes disinterestedness as a basic institutional element. […] A passion for knowledge, idle curiosity, altruistic concern with the benefit to humanity […] have been attributed to the scientist." (pages 275–276 of ref. [14]).

(4) Organized scepticism. This "involves a latent questioning of certain bases of established routine, authority, vested procedures and the realm of the 'sacred' generally. […] Science which asks questions of fact concerning every phase of nature and society comes into psychological, not logical, conflict with other attitudes toward these same data which have been crystallized and frequently ritualized by other institutions. Most institutions demand unqualified faith; but the institution of science makes scepticism a virtue." (pages 264–265 of ref. [14]).

In general, when Mertonian norms are carried over to the field of statistics, general themes include the need to be transparent, to acknowledge uncertainty and to be open to alternative interpretations. As such, the Mertonian norms, although proposed over half a century ago, embody the current aspirations to increase the transparency and reproducibility of science. Critically, the principles behind the Mertonian norms can be translated into concrete statistical practices. A non-exhaustive list of these practices include (1) visualizing data; (2) quantifying inferential uncertainty; (3) assessing data preprocessing choices; (4) reporting multiple models; (5) involving multiple analysts; (6) interpreting results modestly; and (7) sharing data and code. We believe that most statisticians would generally endorse these practices[85], barring reasonable exceptions (that is, privacy concerns, severe restrictions of time and money). Here we explain these practices in more detail, including their benefits, limitations and guidelines.

1854. To trace the source of the outbreak, John Snow created a dot map that displayed the homes of the deceased as well as the water pumps in the neighbourhood (Fig. 1). The scatter of the data showed that the deaths clustered around a particular water pump in Broad Street, suggesting that the disease was waterborne instead of airborne[22]. In response to John Snow's request, the pump was disabled by removing its handle, which immediately ended the neighbourhood epidemic. It was later discovered that the well belonging to the pump was contaminated with sewage, which caused the outbreak in the neighbourhood.

For an example of how data visualization can reveal model misspecification, consider Anscombe's quartet[23] shown in Fig. 2. The four scatter plots all have identical summary statistics (that is, mean, standard deviation and Pearson correlation coefficient values). By visually inspecting the panels, it becomes obvious that the bivariate relation is fundamentally different for each panel[24].

Current status. Since William Playfair (1759–1823) invented the first statistical graphs—such as line graphs and bar charts[25]—data visualization has become an essential part of science. Today, graphs are part of most statistical software packages and have become an indispensable tool to perform certain analyses (that is, principal component analysis or prior and posterior predictive checks) or for handling big datasets (that is, through cluster analysis[26]). Technology now enables us to go beyond static visualizations and display the dynamic aspects of the data, for example, using the software packages R Shiny[27] or iNZight[28].

Limitations. Despite the obvious benefits, data visualization also offers the opportunity to mislead, for example, when displaying spurious patterns by either expanding the scale to minimize variation, or by minimizing the scale to accentuate differences[29–31].

Furthermore, the informativeness of a graph often depends on the design capabilities of the researcher and how much thought they put into what information should be communicated. Scientists without programming experience often find themselves constrained by the options offered in standard graphics software. However, the example of Anscombe's quartet shows that even the simplest plots can be highly informative.

Guidelines. There are no uniform guidelines as to when and which graphical representations should be used. However, there is a fundamental principle of good statistical graphics according to Tufte (page 92 of ref. [32]): "Above all else show the data" (that is, minimize non-data elements). In general, scientists should aim to create a graph that is as clean, informative and as complete as possible. These characteristics are also emphasized in the ASA Ethical Guidelines[33]. The guidelines mention that, to ensure the integrity of data and methods, the ethical statistician "[i]n publications and reports, conveys the findings in ways that are both honest and meaningful to the user/reader. This includes tables, models, and graphics" (page 3 of ref. [33]).

Beyond that, guidelines depend on the individual aspects of the data (that is, the complexity of the data and experimental design) and context (compare with ref. [34]); here we refer the interested reader to the numerous manuals describing good practices in graphical representation of statistical information[32,35–40].

Quantifying inferential uncertainty

Description. By reporting the precision with which model parameters are estimated, the analyst communicates the inevitable uncertainty that accompanies any inference from a finite sample.

Benefits and example. Only by assessing and reporting inferential uncertainty is it possible to make any claim about the degree to which results from the sample generalize to the population. For example, Strack et al.[41] studied whether participants rate cartoons to be funnier when they hold a pen with their teeth (which induces a smile) instead of holding it with their lips (which induces a pout). On a ten-point Likert scale, the authors observed a raw effect size of 0.82 units. For the interpretation of this result, it is essential to know the associated inferential uncertainty. In this case, the 95% confidence interval ranges from −0.05 to 1.69, indicating that the data are not inconsistent with a large range of effect-size estimates (including effect sizes that are negligible or negative).

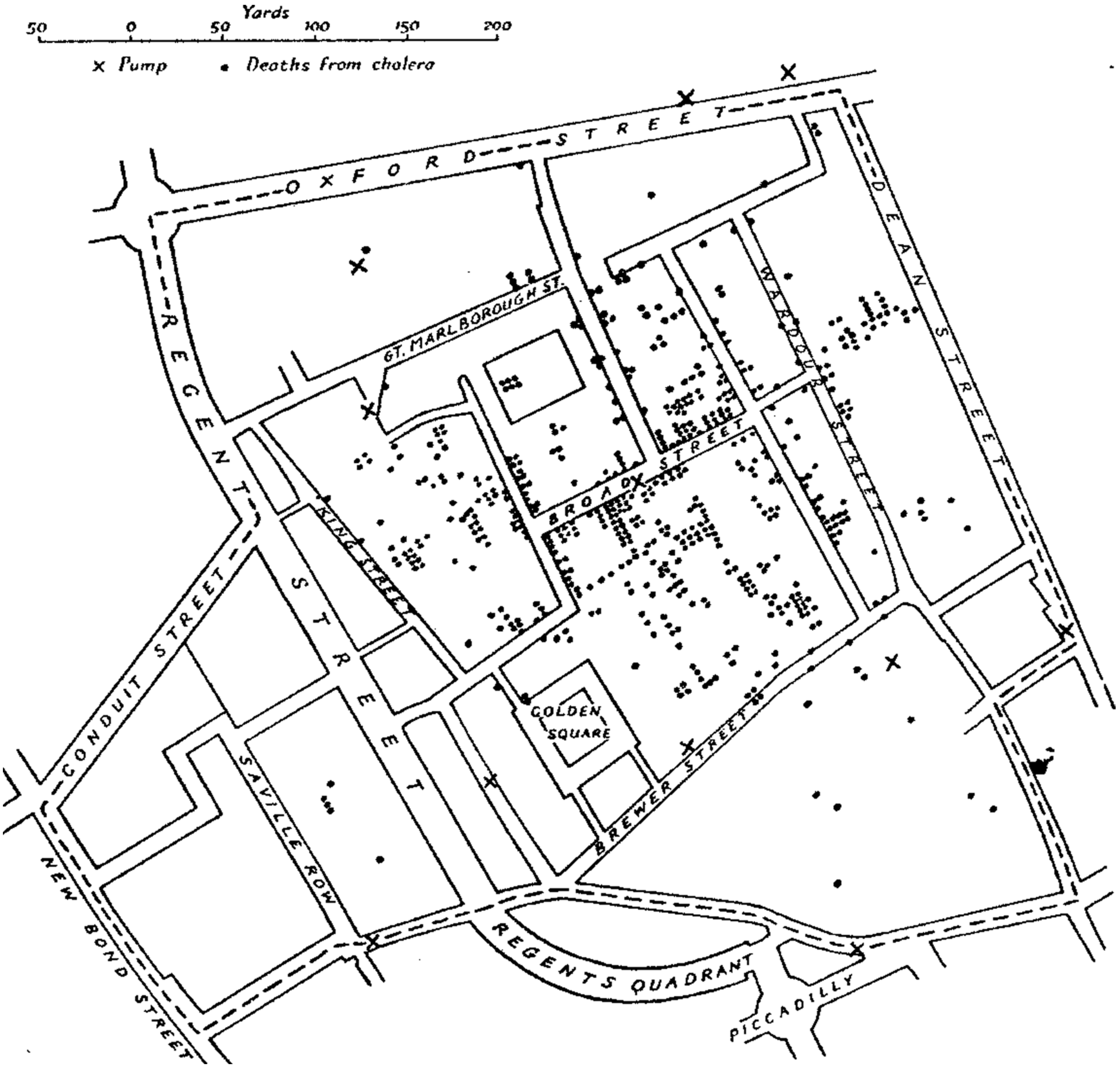

Fig. 1 | Recreation of John Snow's map of the distribution of deaths from cholera. The points represent homes of the deceased individuals and the crosses represent water pumps. The contaminated water pump that triggered the cholera epidemic is located on Broad Street. Reproduced with permission from *The Geographical Journal*[22].

Current status. In almost all statistics courses, students are taught to provide not only the summary of statistical tests (such as *F*, *t* and *P* values and the associated degrees of freedom), but also parameter point-estimates (that is, regression weights, effect sizes) and their associated uncertainty (that is, standard error, posterior distribution, confidence intervals, credible intervals). Nevertheless, there exists a gap between what is taught and what is practiced. Studies of published articles in physiology[20], the social sciences[42] and medicine[43,44] revealed that error bars, standard errors or confidence intervals were not always presented. Popular metrics such as Cronbach's alpha (a measure of test score reliability) are also rarely presented with a measure of inferential uncertainty.

Limitations. We believe that there are no acceptable excuses for omitting a measure of inferential uncertainty in any report.

Although not a limitation as such, it should be noted that inferential uncertainty always needs to be quantified relative to the inferential goal—does a researcher want to generalize across people, stimuli, time points or another dimension? The proper way of computing standard errors depends on the researcher's purpose.

Guidelines. Various guidelines strongly recommend that effect-size estimates are accompanied by measures of uncertainty in the form of standard errors or confidence intervals. For example, the publication manual of the American Psychological Association (6th edn, page 34) states: "When point estimates (that is, sample means or regression coefficients) are provided, always include an associated measure of variability (precision), with an indication of the specific measure used (that is, the standard error)". Furthermore, the International Committee of Medical Journal Editors[45] explicitly recommend to "[w]hen possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals)" (page 17 of ref. [45]).

Assessing data preprocessing choices

Description. By assessing the impact of plausible alternative data preprocessing choices (that is, examining the 'data multiverse'[46]), the analyst determines the extent to which the finding under scrutiny is either fragile or sturdy.

Benefits and example. A data multiverse analysis reveals the fragility or sturdiness of the finding under plausible alternative data

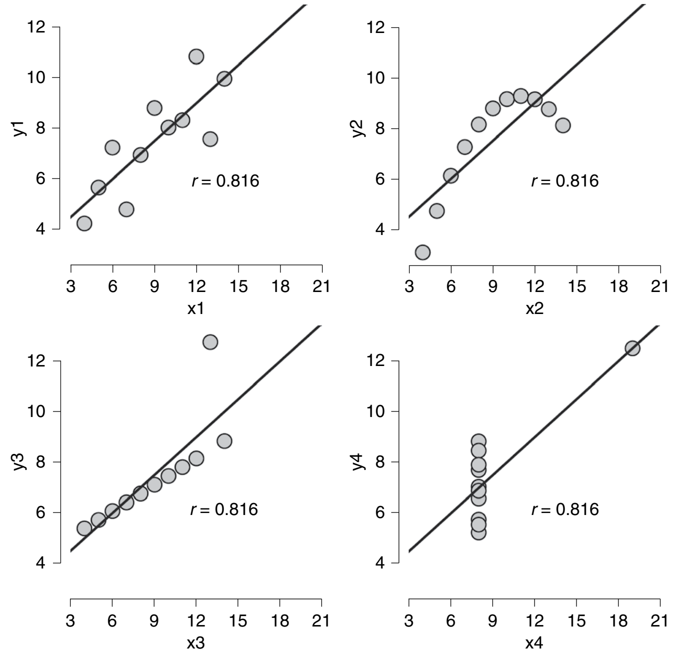

Fig. 2 | Anscombe's quartet emphasizes the importance of data visualization to detect model misspecification. Although the four datasets are equivalent in terms of their summary statistics, the Pearson correlation is appropriate only for the dataset in the top left panel. Figure available at https://www.shinyapps.org/apps/RGraphCompendium (CC BY).

preprocessing choices. This prevents researchers from falling prey to hindsight bias and motivated reasoning, which may lead them to unwittingly report only the preprocessing pipeline that yields the most compelling result[47,48]. But even a completely unbiased analysis will benefit from a data multiverse analysis, as it reveals uncertainty that would otherwise remain hidden.

For example, Steegen et al.[46] re-examined the results of Durante et al.[49], who reported an interaction between relationship status (that is, single or not) and menstrual cycle (that is, fertile or not) on reported religiosity. After applying a series of 180 different data preprocessing procedures (that is, five different ways to split women into high versus low fertility), the multiverse reanalysis showed that the resulting 180 P values were distributed uniformly between 0 and 1, indicating that the reported interaction is highly fragile.

Current status. The idea of assessing sensitivity to data-preprocessing choices dates back at least to De Groot (page 190 of ref. [47]) and Leamer (page 308 of ref. [50]) and was revived by Simmons et al.[48] and by Steegen et al.[46]. In the field of functional magnetic resonance imaging, both Carp[51] and Poldrack et al.[52] emphasized the hidden influence of different plausible preprocessing pipelines. In psychology, recent applications are Bastiaansen et al.[3] and Wessel et al.[53]. Nevertheless, the overwhelming majority of empirical articles does not report the results of a data multiverse analysis.

Limitations. A pragmatic limitation of the data multiverse lies in the extra work that it entails. Another limitation can be found in ambiguities surrounding the definition of the data multiverse. The analyst has to determine what constitutes a sufficiently representative set of preprocessing choices and whether all preprocessing choices are equally plausible, such that they should be given equal weight in the multiverse analysis. A final limitation is that it is not always clear how to interpret the results of a data multiverse analysis. Interpretation can be facilitated with certain graphical formats that cluster related pipelines (that is, specification curves)[54].

Guidelines. Some specific guidelines on assessing data preprocessing choices are offered by Simmons et al.[48] (see the requirements for authors, numbers 5 and 6), but it is difficult to provide general guidelines as "[…] a multiverse analysis is highly context-specific and inherently subjective. Listing the alternative options for data construction requires judgement about which options can be considered reasonable and will typically depend on the experimental design, the research question, and the researchers performing the research" (page 709 of ref. [46]). More general guidelines that relate exclusively to the reporting of preprocessing choices are given in the ASA Ethical Guidelines[33]. These mention that, to ensure the integrity of data and methods, the ethical statistician "[w]hen reporting on the validity of data used, acknowledges data editing procedures, including any imputation and missing data mechanisms" (page 2 of ref. [33]).

Reporting multiple models

Description. By assessing the impact of plausible alternative statistical models (that is, examining the model multiverse), the analyst gauges the extent to which a statistical conclusion is either fragile or sturdy.

Benefits and example. Similar to the data multiverse analysis discussion in the previous section, a model multiverse analysis examines the fragility or sturdiness of the finding under plausible alternative statistical modelling choices. Modelling choices comprise differences in estimators and fitting regimes, but also in model specification and variable selection. Reporting the outcomes of multiple plausible models reveals uncertainty that would remain hidden if only a single model were entertained. Moreover, this practice protects analysts against hindsight bias and motivated reasoning, which may unwittingly lead them to select the single model that produces the most flattering conclusion. For example, Patel et al.[55] quantified the variability of results under different model specifications. They considered 13 clinical, environmental and physiological variables as potential covariates for the association of 417 self-reported, clinical and molecular phenotypes with all-cause mortality. As a consequence, they computed P values for $2^{13}=8,192$ models and examined the instability of the inference, which they call the 'vibration of effects'.

Current status. Although the idea of the model multiverse dates back at least to De Groot[47] and Leamer[50], most empirical researchers still base their conclusion on only a single analysis (but also see refs. [56,57]).

Limitations. As was the case for the construction of the data multiverse, a pragmatic limitation of the model multiverse lies in the extra work that it entails for the analyst as well as the reader. Recent work suggests that the number of plausible models can be very large[4,7]. Multiverses also vary in their informativeness, and readers need to assess themselves whether a multiverse features notably distinct models or just runs the essentially same model multiple times. Model spaces can be overwhelming; any single analyst will naturally be drawn towards the subset of models that they are familiar with (or, unwittingly, the subset of models that yields the result that is most flattering or most in line with prior expectations). Moreover, Del Giudice et al. (page 5 of ref. [58]) argue that "By inflating the size of the analysis space, the combinatorial explosion of unjustified specifications may, ironically, exaggerate the perceived exhaustiveness and authoritativeness of the multiverse while greatly reducing the informative fraction of the multiverse. At the same time, the size of the specification space can make it harder to inspect the results for potentially relevant findings. If unchecked, multiverse-style analyses can generate analytic 'black holes': Massive analyses that swallow true effects of interest but, due to their perceived exhaustiveness

Box 2 | Seven Mertonian statistical procedures

This box outlines how each of the seven procedures discussed in this Perspective fulfil the Mertonian norms. An overview is given in the table below.

	Communalism	Universalism	Disinterestedness	Organized scepticism
(1) Visualizing data	Yes		Yes	Yes
(2) Quantifying inferential uncertainty	Yes		Yes	Yes
(3) Assessing data preprocessing choices	Yes		Yes	Yes
(4) Reporting multiple models	Yes		Yes	Yes
(5) Involving multiple analysts		Yes	Yes	Yes
(6) Interpreting results modestly			Yes	Yes
(7) Sharing data and code	Yes	Yes	Yes	Yes

(1) Visualizing data. Well-designed visualizations show at a glance the key aspects of the data. Moreover, by giving the reader a more complete picture of the data and related statistics, visualizations can either support or weaken a conclusion drawn by the researcher, or help the reader to find alternative ways of interpreting the results and analysing the data.

(2) Quantifying inferential uncertainty. Acknowledging inferential uncertainty (that is, by presenting standard errors or confidence intervals) contributes to open communication. Moreover, quantifying inferential uncertainty signals that researchers are openly acknowledging the extent to which their measurements are imprecise, especially when the sample size is small. Finally, explicitly acknowledging inferential uncertainty may prompt readers to question how well the results from the sample generalize to the population.

(3) Assessing data preprocessing choices. When researchers share the results from only a single data preprocessing pipeline, they may unintentionally hide important information. If a result proves sensitive to particular preprocessing choices, this warrants scepticism and may initiate a debate on the importance and plausibility of relevant data preprocessing choices (compare with page 308 of ref. [50]).

(4) Reporting multiple models. Similar to the previous section, reporting results from only a single model may unintentionally hide important information.

(5) Involving multiple analysts. The multiple-analysts approach can reveal whether different teams of analysts reach converging or diverging conclusions from the same dataset. By including other analysts with different backgrounds and interests, the potential impact of self-interest of any single analyst is counteracted. The multiple-analysts approach also stimulates scepticism by bringing to light alternative statistical perspectives on the data.

(6) Interpreting results modestly. Disinterested analysts arguably have little need to exaggerate claims, impress reviewers and downplay signs of model misfit. Analysts who facilitate organized scepticism do not attempt to suppress doubt—they are not defensive, and they do not wish to protect their work against good-faith scrutiny from their peers.

(7) Sharing data and code. All secrecy about data is a limitation to knowledge accumulation and violates the ethos of science. All interested researchers should have access to relevant, properly anonymized data. Importantly, sharing data enables sceptical eyes to scrutinize the results, promoting quality control.

and sheer size, trap whatever information is present in impenetrable displays and summaries."

Guidelines. Because the construction of the model multiverse depends on the knowledge and expertise of the analyst, it is challenging to provide general guidelines. However, for relatively simple regression models, clear guidelines do exist[55,59]. Furthermore, Simonsohn et al.[54] suggested a specification curve analysis, and Dragicevic et al.[60] suggest interactive ways of presenting the results. The ASA Ethical Guidelines[33] mention that, to meet the responsibilities towards funders and clients, the ethical statistician "[t]o the extent possible, presents a client or employer with choices among valid alternative statistical approaches that may vary in scope, cost, or precision" (page 3 of ref. [33]). However, the ASA does not mention that researchers share the same responsibility towards their scientific colleagues, although this may be implicit.

One general recommendation for constructing a comprehensive model multiverse is to collaborate with statisticians who have complementary expertise, bringing us to the next section.

Involving multiple analysts

Description. By having multiple analysts independently analyse the same dataset, the researcher can decrease the impact of analyst-specific choices regarding data preprocessing and statistical modelling.

Benefits and example. The multiple-analysts approach reveals the uncertainty that is due to the subjective choices of a single analyst and promotes the application of a wider range of statistical techniques. When the conclusions of the analysts converge, this bolsters one's confidence that the finding is robust; when the conclusions diverge, this undercuts that confidence and stimulates a closer look at the statistical reasons for the lack of consensus.

The multiple-analysts approach was used, for example, in a study by Silberzahn et al.[7], in which 29 teams of analysts examined, using the same dataset, whether the skin tone of soccer players influences their probability of getting a red card. Although most of the analysis teams reported that players with a darker skin tone have a higher probability of getting a red card, some of the teams reported null results. The analysis approach used by the teams differed widely, both with respect to data preprocessing and statistical modelling (that is, included covariates, link functions, assumption of hierarchical structure).

Current status. A precursor to the multiple-analysts approach concerns the 1857 'Cuneiform competition', in which four scholars independently translated a previously unseen ancient Assyrian inscription[61]. The overlap between their translations—which were sent to the Royal Asian Society in sealed envelopes, and were simultaneously opened and inspected by a separate committee of examiners—was striking and put to rest any doubts concerning the method used to decipher such inscriptions. The multiple-analysts approach never caught on in practice, although recent examples exist in psychology and neuroscience[3–5,7,62,63].

Limitations. As was the case for the construction of the data multiverse and the model multiverse, a pragmatic limitation of the multiple-analyst approach lies in the extra work that it entails, specifically with respect to (1) finding knowledgeable analysts who are interested in participating; (2) documenting the dataset, describing the research question and identifying the target of statistical inference; (3) collating the initial responses from each team and potentially coordinating a review and feedback round. Although differences in opinion should be respected, there needs to be ways to filter out analysis approaches that involve clear mistakes. An additional limitation concerns possible homogeneity of the analysts. For example, all analysts involved could be rigidly educated in the same school of thought, share cultural or social biases or just make the same mistake. In such a case, the results may create an inflated sense of certainty in the conclusion that was reached. These potential limitations can be mitigated by selecting a diverse group of analysts and incorporating feedback and revision options in the process[7], a round-table discussion[5] or, more systematically, a Delphi approach[64].

Guidelines. There are no explicit guidelines concerning the multiple-analysts approach. We propose that the optimal number of analysts to be included depends on factors such as the complexity of the data, the importance of the research question (that is, a clinical trial on the effectiveness of a new drug against coronavirus disease 2019 (COVID-19) warrants a relatively large number of analysts) and the probability that the analysts could reasonably reach a different conclusion (that is, there may be multiple ways to interpret the research question, and there may be multiple dependent variables and predictor variables that could or could not be relevant).

When analysts are selected, care should be taken to ensure heterogeneity, diversity and balance. Specifically, one should be mindful of the potential biasing effects of specific background knowledge, culture, education and career stage of the analyst.

The ASA Guidelines emphasize the legitimacy and value in alternative analytic approaches, stating that "[t]he practice of statistics requires consideration of the entire range of possible explanations for observed phenomena, and distinct observers […] can arrive at different and potentially diverging judgements about the plausibility of different explanations" (page 5 of ref. [33]).

Interpreting results modestly

Description. By modestly interpreting the results, the analyst explicitly acknowledges any remaining doubts concerning the importance, replicability and generalizability of the scientific claims at hand.

Benefits and example. Modestly presented scientific claims enable the reader to evaluate the outcomes for what they usually are: not final, but tentative results pointing in a certain direction, with considerable uncertainty surrounding their generalizability and scope. Overselling results might lead to the misallocation of public resources towards approaches that are in fact not properly validated and not ready for application in practice. Furthermore, researchers themselves risk losing long-term credibility for short-term gains of greater attention and higher citation counts. Moreover, after having publicly committed to a bold claim, it becomes difficult to admit that one's initial assessment was wrong; in other words, overconfidence is not conducive to scientific learning.

Scientists of true modesty remain doubtful even at moments of great success. For example, when James Chadwick found experimental proof of neutrons, the discovery that earned him the Nobel prize, he communicated it modestly under the title 'Possible existence of neutron'[65].

Current status. Tukey[66] already remarked that "Laying aside unethical practices, one of the most dangerous [[…] practices of data analysis […]] is the use of formal data-analytical procedures for sanctification, for the preservation of conclusions from all criticism, for the granting of an imprimatur." (page 13 of ref. [66]). Almost 60 years later, an editorial in *Nature Human Behaviour* warned its readers about "conclusive narratives that leave no room for ambiguity or for conflicting or inconclusive results" (page 1 of ref. [67]). Similarly, Simons et al.[68] suggested adding a mandatory 'constraints on generality' statement in the discussion section of all primary research articles in the field of psychology to prevent authors from making wildly exaggerated claims of generality. This suggests that scientific modesty is rarer than we would expect if Mertonian norms were widely adopted. There are some clear indications of a lack of modesty. First, the frequency of stronger language (words such as amazing, ground-breaking, unprecedented) seemed to have increased in the past few decades[69]. Second, dichotomization of findings (that is, ignoring the uncertainty inherent to statistical inference) is common practice (ref. [42], see paragraph 4.3). Third, textbooks (which are typically a reflection of current practice) on how to write papers often explicitly encourage authors to overclaim[70,71].

Limitations. Publications and grants are important for scientific survival. Coupled with the fact that journals and funders often prefer ground-breaking and unequivocal outcomes, it may be detrimental to one's success to modestly interpret the results. The encouragement of this Mertonian practice may require change at an institutional level, although some have argued that scientists should not hide behind the system when defending their behaviour[72].

Guidelines. There are several ways that we can contribute to increasing intellectual modesty. First, we could encourage intellectual modesty in the work of others when we act as reviewers of papers and grant proposals[73]. As a reviewer's career is independent of how they evaluate a paper, they can make a positive review conditional on a more modest presentation of outcomes. Hoekstra and Vazire[73] present a list of suggestions for increasing modesty in the traditional sections of an empirical article, which can be used by authors as well. One example includes "Titles should not state or imply stronger claims than are justified (that is, causal claims without strong evidence)" (page 16 of ref. [73]) .

The ASA Guidelines also state: "[t]he ethical statistician is candid about any known or suspected limitations, defects, or biases in the data that may affect the integrity or reliability of the statistical analysis" (page 2 of ref. [33]).

Sharing data and code

Description. By sharing data and analysis code, researchers provide the basis for their scientific claims. Ideally, data and code should be shared publicly, freely and in a manner that facilitates reuse.

Benefits and example. As there are many different ways of processing and analysing data[7,46], sharing code promotes reproducibility and encourages sensitivity analyses. Sharing data and code also enables other researchers to establish the validity of the original analyses, it can facilitate collaboration, but it can also serve as protection against data loss. When publishing his theory on general intelligence, Spearman[74] shared his data as an appendix to the article. A century later, this act of foresight enabled scientists to use this dataset for both research and education. Because Spearman made his data publicly available, other researchers could establish the reproducibility and generalizability of the findings.

Current status. Data sharing has never been easier. Public repositories offer free storage space for research materials, data (that is, the Open Science Framework) and code (that is, GitHub). While data sharing is not yet a general practice in most scientific fields, several recent initiatives (that is, Open Data/Code/Materials badges[75]), standards (TOP Guidelines[76]), journals (that is, *Scientific Data*) and checklists (that is, Transparency Checklist[77]) are helping to promote this research practice. When sharing raw data is unfeasible, researchers can make aggregated data summaries available—for example, the data used to generate certain plots or covariance matrices of involved variables.

Limitations. Restrictions imposed by funders, ethics review boards in universities and other institutions, collaborators and legal contracts may limit the extent to which data can be publicly shared. There may also be practical considerations (that is, sharing big data), data use agreements, privacy rights and institutional policies that can curtail sharing intentions. What remains central is to inform the readers about the accessibility of the data of the analysis. Note that these limitations should not apply to the analysis code provided that code is solely reflective of the researcher's analysis actions and is free of any data privacy issues.

Guidelines. An important principle of sharing data is that they should be findable, accessible, interoperable and reusable (FAIR[78]). Several guides are available discussing the practical[79] and ethical[80] aspects of data sharing. Researchers should follow the data sharing procedures and requirements of their fields[81,82] and indicate the accessibility of the data in the research report[76,83]. The ASA Ethical Guidelines[33] for Statistical Practice state that the ethical statistician "[p]romotes sharing of data and methods as much as possible", and "[m]akes documentation suitable for replicate analyses, metadata studies, and other research by qualified investigators." (page 5 of ref. [33]).

Concluding comments

If the statistical literature is any guide, one may conclude that statisticians rarely agree with each other. For example, the 2019 special issue in *The American Statistician* featured 43 articles on *P* values and, in their editorial, Wasserstein et al.[13] stated that "the voices in the 43 papers in this issue do not sing as one". However, despite the continuing disagreements about the foundations of statistical inference, we believe that there is nevertheless much common ground among statisticians, specifically with respect to the ethical aspects of their profession. To examine this ethical dimension more systematically, we started by considering the Mertonian norms that characterize the ethos of science and outlined a non-exhaustive list of seven concrete, teachable, and implementable practices that we believe need wider propagation.

In essence, these practices are about promoting transparency and the open acknowledgement of uncertainty. With agreement on such practices explicitly acknowledged, we believe that commonly discussed contentious issues (that is, *P* values) may become less crucial. Indeed, in a letter to his frequentist nemesis Sir Ronald Fisher, the arch-Bayesian Sir Harold Jeffreys wrote "Your letter confirms my previous impression that it would only be once in a blue moon that we would disagree about the inference to be drawn in any particular case, and that in the exceptional cases we would both be a bit doubtful" (page 162 of ref. [84]).

We hope that the proposed statistical practices will improve the quality of data analysis across the board, especially in applied disciplines that are perhaps unfamiliar with the ethical aspects of statistics, aspects that a statistician may take for granted. Furthermore, instead of counting on them to be absorbed through osmosis, we believe that it is important to include these ethical considerations—and their statistical consequences—explicitly in the statistics curricula. Statistical techniques other than those discussed here may also further the Mertonian ideals. We hope that this contribution provides the impetus for a deeper exploration of how data analysis in applied fields can become more transparent, more informative and more open about the uncertainties that inevitably arise in any statistical data analysis problem.

Received: 2 March 2021; Accepted: 6 September 2021;
Published online: 11 November 2021

References

1. Good, I. J. 46656 varieties of Bayesians. *Am. Stat.* **25**, 62–63 (1971).
2. Aczel, B. et al. Discussion points for Bayesian inference. *Nat. Hum. Behav.* **4**, 561–566 (2020).
3. Bastiaansen, J. A. et al. Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *J. Psychosom. Res.* **137**, 110211 (2020).
4. Botvinik–Nezer, R. et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020).
5. van Dongen, N. et al. Multiple perspectives on inference for two simple statistical scenarios. *Am. Stat.* **73**, 328–339 (2019).
6. Salganik, M. J. et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc. Natl Acad. Sci. USA* **117**, 8398–8403 (2020).
7. Silberzahn, R. et al. Many analysts, one data set: making transparent how variations in analytic choices affect results. *Adv. Methods Pract. Psychol. Sci.* **1**, 337–356 (2018).
8. Amrhein, V., Greenland, S. & McShane, B. B. Retire statistical significance. *Nature* **567**, 305–307 (2019).
9. Benjamin, D. J. et al. Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10 (2018).
10. Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (eds) *What if There Were No Significance Tests?* (Lawrence Erlbaum, Mahwah, 1997).
11. McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. Abandon statistical significance. *Am. Stat.* **73**, 235–245 (2019).
12. Wasserstein, R. L. & Lazar, N. A. The ASA's statement on p–values: context, process, and purpose. *Am. Stat.* **70**, 129–133 (2016).
13. Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. Moving to a world beyond '$p<0.05$'. *Am. Stat.* **73**, 1–19 (2019).
14. Merton, R. K. (ed.) in *The Sociology of Science: Theoretical and Empirical Investigations* 267–278 (Univ. of Chicago Press, 1973).
15. Tukey, J. W. *Explanatory Data Analysis* (Addison–Wesley, 1977).
16. Gelman, A. Exploratory data analysis for complex models. *J. Comput. Graph. Stat.* **13**, 755–779 (2004).
17. Gabry, J., Simpson, D., Vehtari, A., Betancourt, M. & Gelman, A. Visualization in Bayesian workflow. *J. R. Stat. Soc. A* **182**, 389–402 (2019).
18. Heathcote, A., Brown, S. D. & Wagenmakers, E.-J. in *An Introduction to Model-Based Cognitive Neuroscience* (eds Forstmann, B. & Wagenmaker, E.-J.) 25–48 (Springer, 2015).
19. Kerman, J., Gelman, A., Zheng, T. & Ding, Y. in *Handbook of Data Visualization* (eds Chen, C. et al.) 709–724 (Springer, 2008).
20. Weissgerber, T. L., Milic, N. M., Winham, S. J. & Garovic, V. D. Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS Biol.* **13**, e1002128 (2015).
21. Healy, K. & Moody, J. Data visualization in sociology. *Ann. Rev. Sociol.* **40**, 105–128 (2014).
22. Gilbert, E. W. Pioneer maps of health and disease in England. *Geogr. J.* **124**, 172–183 (1958).
23. Anscombe, F. J. Graphs in statistical analysis. *Am. Stat.* **27**, 17–21 (1973).
24. Matejka, J. & Fitzmaurice, G. Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proc. 2017 CHI Conference on Human Factors in Computing Systems* 1290–1294 (2017).
25. Playfair, W. *Commercial and Political Atlas: Representing, by Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the Whole of the Eighteenth Century* (1786).
26. Everitt, B. S., Landau, S., Leese, M. & Stahl, D. *Cluster Analysis* (John Wiley & Sons, 2011).
27. Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. shiny: web application framework for R, version 1.7.0, http://CRAN.R-project.org/package=shiny (2020).
28. iNZight Team iNZight v.4.0.2. https://inzight.nz (2020).
29. Cairo, A. *How Charts Lie: Getting Smarter about Visual Information* (WW Norton & Company, 2019).
30. Gelman, A. Why tables are really much better than graphs. *J. Comput. Graph. Stat.* **20**, 3–7 (2011).
31. Wainer, H. How to display data badly. *Am. Stat.* **38**, 137–147 (1984).

32. Tufte, E. R. *The Visual Display of Quantitative Information* (Graphics Press, 1973).
33. Committee on Professional Ethics of the American Statistical Association *Ethical Guidelines for Statistical Practice,* https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx (2018).
34. Diamond, L. & Lerch, F. J. Fading frames: data presentation and framing effects. *Decis. Sci.* **23**, 1050–1071 (1992).
35. Chen, C., Härdle, W. & Unwin, A. (eds) *Handbook of Data Visualization* (Springer, 2008).
36. Cleveland, W. S. & McGill, R. Graphical perception: theory, experimentation, and application to the development of graphical methods. *J. Am. Stat. Assoc.* **79**, 531–554 (1984).
37. Gelman, A., Pasarica, C. & Dodhia, R. Let's practice what we preach: turning tables into graphs. *Am. Stat.* **56**, 121–130 (2002).
38. Mazza, R. *Introduction to Information Visualization* (Springer Science & Business Media, 2009).
39. Wilke, C. O. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures* (O'Reilly Media, 2019).
40. Wilkinson, L. *The Grammar of Graphics* (Springer Science & Business Media, 1999).
41. Strack, F., Martin, L. L. & Stepper, S. Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *J. Pers. Soc. Psychol.* **54**, 768–777 (1988).
42. Hoekstra, R., Finch, S., Kiers, H. A. & Johnson, A. Probability as certainty: dichotomous thinking and the misuse of *p*-values. *Psychon. Bull. Rev.* **13**, 1033–1037 (2006).
43. Cooper, R. J., Schriger, D. L. & Close, R. J. Graphical literacy: the quality of graphs in a large-circulation journal. *Ann. Emerg. Med.* **40**, 317–322 (2002).
44. Schriger, D. L., Sinha, R., Schroter, S., Liu, P. Y. & Altman, D. G. From submission to publication: a retrospective review of the tables and figures in a cohort of randomized controlled trials submitted to the British Medical Journal. *Ann. Emerg. Med.* **48**, 750–756 (2006).
45. International Committee of Medical Journal Editors *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals*, http://www.icmje.org/icmje-recommendations.pdf (2019).
46. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.* **11**, 702–712 (2016).
47. De Groot, A. D. The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas]. *Acta Psychologica* **148**, 188–194 (2014).
48. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False–positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
49. Durante, K. M., Rae, A. & Griskevicius, V. The fluctuating female vote: politics, religion, and the ovulatory cycle. *Psychol. Sci.* **24**, 1007–1016 (2013).
50. Leamer, E. E. Sensitivity analyses would help. *Am. Econ. Rev.* **75**, 308–313 (1985).
51. Carp, J. On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. *Front. Neurosci.* **6**, 149 (2012).
52. Poldrack, R. A. et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* **18**, 115–126 (2017).
53. Wessel, I., Albers, C., Zandstra, A. R. E. & Heininga, V. E. A multiverse analysis of early attempts to replicate memory suppression with the Think/No-think Task. *Memory* **28**, 870–887 (2020).
54. Simonsohn, U., Nelson, L. D. & Simmons, J. P. Specification curve analysis. *Nat. Hum. Behav.* **4**, 1208–1214 (2020).
55. Patel, C. J., Burford, B. & Ioannidis, J. P. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J. Clin. Epidemiol.* **68**, 1046–1058 (2015).
56. Athey, S. & Imbens, G. W. Machine learning methods that economists should know about. *Ann. Rev. Econ.* **11**, 685–725 (2019).
57. Levine, R. & Renelt, D. A sensitivity analysis of cross-country growth regressions. *Am. Econ. Rev.* **82**, 942–963 (1992).
58. Del Giudice, M., Gangestad, S. W. & Steven, W. A traveler's guide to the multiverse: promises, pitfalls, and a framework for the evaluation of analytic decisions. *Adv. Methods Pract. Psychol. Sci.* **4**, 1–15 (2021).
59. Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. Bayesian model averaging: a tutorial. *Stat. Sci.* **14**, 382–401 (1999).
60. Dragicevic, P., Jansen, Y., Sarma, A., Kay, M. & Chevalier, F. Increasing the transparency of research papers with explorable multiverse analyses. In *Proc. 2019 CHI Conference on Human Factors in Computing Systems* 1–15 (2019).
61. Rawlinson, H., Talbot, F., Hincks, E. & Oppert, J. *Inscription of Tiglath Pileser I., king of Assyria, B.C. 1150, as translated by Sir Henry Rawlinson, Fox Talbot, Esq., Dr. Hincks, and Dr. Oppert (Published by the Royal Asiatic Society)* (J. W. Parker and Son, 1857).
62. Boehm, U., Hawkins, G. E., Brown, S. D., van Rijn, H. & Wagenmakers, E.-J. Of monkeys and men: impatience in perceptual decision–making. *Psychon. Bull. Rev.* **23**, 738–749 (2016).
63. Dutilh, G. et al. The quality of response time data inference: a blinded, collaborative assessment of the validity of cognitive models. *Psychon. Bull. Rev.* **26**, 1051–1069 (2019).
64. Thangaratinam, S. & Redman, C. W. The delphi technique. *Obstet. Gynaecol.* **7**, 120–125 (2005).
65. Chadwick, J. Possible existence of a neutron. *Nature* **129**, 312 (1932).
66. Tukey, J. W. The future of data analysis. *Ann. Math. Stat.* **33**, 1–67 (1962).
67. Tell it like it is. *Nat. Hum. Behav.* **4**, 1 (2020).
68. Simons, D. J., Shoda, Y. & Lindsay, D. S. Constraints on generality (cog): a proposed addition to all empirical papers. *Perspect. Psychol. Sci.* **12**, 1123–1128 (2017).
69. Vinkers, C. H., Tijdink, J. K. & Otte, W. M. Use of positive and negative words in scientific pubmed abstracts between 1974 and 2014: Retrospective analysis. *BMJ* **351**, h6467 (2015).
70. Bem, D. J. in *The Compleat Academic: A Practical Guide for the Beginning Social Scientist* (eds Zanna, M. R. & Darley, J. M.) 171–201 (Lawrence Erlbaum Associates, 1987).
71. van Doorn, J. et al. Strong public claims may not reflect researchers' private convictions. *Significance* **18**, 44–45 (2021).
72. Yarkoni, T. No, it's not the incentives–it's you, https://www.talyarkoni.org/blog/2018/10/02/no-its-not-the-incentives-its-you/ (2018).
73. Hoekstra, R. & Vazire, S. Aspiring to greater intellectual humility in science. Preprint at https://doi.org/10.31234/osf.io/edh2s (2020).
74. Spearman, C. General intelligence, objectively determined and measured. *Am. J. Psychol.* **15**, 201–293 (1904).
75. Kidwell, M. C. et al. Badges to acknowledge open practices: a simple, low cost, effective method for increasing transparency. *PLoS Biol.* **14**, e1002456 (2016).
76. Nosek, B. et al. Promoting an open research culture. *Science* **348**, 1422–1425 (2015).
77. Aczel, B. et al. A consensus-based transparency checklist. *Nat. Hum. Behav.* **4**, 4–6 (2020).
78. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
79. Klein, O. et al. A practical guide for transparency in psychological science. *Collabra Psychol.* **4**, 20 (2018).
80. Alter, G. & Gonzalez, R. Responsible practices for data sharing. *Am. Psychol.* **73**, 146–156 (2018).
81. Wagenmakers, E.-J., Kucharsky, S. & The JASP Team (eds) *The JASP Data Library* (JASP Publishing, 2020).
82. Taichman, D. B. et al. Data sharing statements for clinical trials: a requirement of the International Committee of Medical Journal Editors. *JAMA* **317**, 2491–2492 (2017).
83. Aalbersberg, I. J. et al. Making science transparent by default; introducing the TOP statement. Preprint at https://osf.io/sm78t (2018).
84. Bennett, J. H. (ed.) *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher* (Clarendon Press, 1990).
85. Anderson, M. S., Martinson, B. C. & De Vries, R. Normative dissonance in science: results from a national survey of US scientists. *J. Empir. Res. Hum. Res. Ethics* **2**, 3–14 (2007).

Acknowledgements

We are grateful to N. Lazar for her comments on a draft version. We thank everyone who was involved in drafting the initial list of statistical procedures during the hackathon that took place at the 2019 meeting of the Society for the Improvement of Psychological Science in Rotterdam, The Netherlands. This work was supported in part by a European Research Council (ERC) grant to E.-J.W. (no. 283876), a Netherlands Organisation for Scientific Research (NWO) grant to A. Sarafoglou (no. 406-17-568), as well as a Dutch scientific organization Vidi grant from the NWO to D.v.R. (no. 016.Vidi.188.001).

Author contributions

Conceptualization: E.-J.W., A. Sarafoglou and B.A. Project administration: B.A. Writing (original draft preparation): E.-J.W., A. Sarafoglou, C.A., J.A., Š.B., N.v.D., R.H., D.M., D.v.R., A. Sluga, J.T. and B.A. Writing (review and editing): E.-J.W., A. Sarafoglou, S.A., C.A., J.A., Š.B., N.v.D., R.H., D.M., D.v.R., A. Sluga, F.S., J.T. and B.A.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Eric-Jan Wagenmakers.

Peer review information *Nature Human Behaviour* thanks David Hand and Ulf Toelch for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.