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]

[1]Department of Psychology, University of Amsterdam, The Netherlands
[2]School for Public Health and Primary Care, Maastricht University, The Netherlands
[3]Heymans Institute of Psychological Research, University of Groningen, The Netherlands
[4]Donders Institute for Brain, Cognition and Behaviour, Radboud University, The Netherlands
[5]Faculty of Business Administration, Prague University of Economics, Czech Republic
[6]Department of Educational Science, University of Groningen, The Netherlands
[7]School of Psychology and Centre for Brain Research, The University of Auckland, New Zealand
[8]Department of Psychology, University of Groningen, The Netherlands
[9]Rotterdam School of Management, Erasmus University Rotterdam, The Netherlands
[10]Department of Psychology, University of Münster, Germany
[11]Office of Research and Academia-Government-Community Collaboration, Hiroshima University, Japan
[12]Institute of Psychology, ELTE Eotvos Lorand University, Hungary

March 2, 2021

[*]Correspondence concerning this article should be addressed to: Eric-Jan Wagenmakers, University of Amsterdam, Nieuwe Achtergracht 129 B, 1018 WT Amsterdam, The Netherlands. E-mail may be sent to ej.wagenmakers@gmail.com.

Abstract

We explore the promise of statistical reform by starting from the assumption that most researchers would endorse Merton's ethos of science as reflected in the four norms of communalism, universalism, disinterestedness, and organized skepticism. Translated to data analysis, these norms imply a need for transparency, a fair acknowledgement of uncertainty, and openness to alternative interpretations. We discuss seven statistical procedures, both old and new, that we believe can positively impact statistical practice in the social and behavioral sciences.

Keywords: Consensus, Transparency, Open Science, Statistical Recommendations

"The collective noun for a group of statisticians is a quarrel."

John Tukey (Brillinger, 2014, p.44)

1 INTRODUCTION

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 vs. 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 is also present within the same school. For instance, 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 an underestimate (Good, 1971; but see Aczel et al., 2020a).

The disagreement also manifests itself in practical application, whenever multiple statisticians and practitioners of statistics find themselves independently analyzing the same data set. Specifically, recent "multiple-analyst" articles show that statisticians rarely use the same analysis, and often draw different conclusions, even for the exact same data set and research question (Bastiaansen et al., 2020; Botvinik–Nezer et al., 2020; van Dongen et al., 2019; Salganik et al., 2020; Silberzahn et al., 2018). Deep disagreement is also exhibited by contradictory guidelines on p-values (e.g., Amrhein et al., 2019; Benjamin et al., 2018; Harlow et al., 1997; McShane et al., 2019; Wasserstein and Lazar, 2016; Wasserstein et al.,

2019). 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, practitioners may be forgiven for adopting a wait-and-see attitude and carrying on as usual.

In this paper, we claim that besides the numerous disputes and outstanding arguments, statisticians share a common set of scientific norms. These norms are almost never articulated explicitly, and our main goal is to bring them to the fore, as we believe that they have considerable relevance for the practice of statistics in the social and behavioural sciences. Here, we focus on the four scientific norms proposed by Merton (1973) (originally published in 1942; see the textbox for a detailed overview of the Mertonian norms), that is, communalism, universalism, disinterestedness, and organized skepticism.

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. Moreover, 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; (7) sharing data and code.[1] We believe that most statisticians would generally endorse these practices, barring reasonable exceptions (e.g., privacy concerns, severe restrictions of time and money). In this paper, we will explain these practices in more detail,

[1]This list was the result of a hackathon that took place at the 2019 meeting of the Society for the Improvement of Psychological Science in Rotterdam, The Netherlands.

including their benefits, limitations and guidelines.

Merton (1973) 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." (Merton, 1973, pp. 273–274)

2. Universalism. "truth-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." (Merton, 1973, p. 270; italics in original)

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." (Merton, 1973, pp. 275-276)

4. Organized Skepticism. 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." (Merton, 1973, p. 264–265; italics in original)

1.1 VISUALIZING DATA

> "Graphs are essential to good statistical analysis."
>
> Frank Anscombe (1973, p. 17)

1.1.1 Description

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

1.1.2 Benefits and Examples

Data visualization is important in all phases of the statistical workflow. In exploratory data analysis, data visualization helps researchers formulate new theories and hypotheses (Tukey, 1977). In model assessment, data visualization supports the detection of model misfit and guides the development of appropriate statistical models (e.g., Gelman, 2004; Gabry et al., 2019; Heathcote et al., 2015; Kerman et al., 2008; Weissgerber et al., 2015). 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 (Healy and Moody, 2014).

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 Dr. John Snow during the cholera outbreak in Soho, London in September 1854. In order to trace the source of the outbreak, Dr. Snow created a dot map that displayed the homes of the deceased as well as the water pumps in the neighborhood (Figure 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

(Gilbert, 1958). Upon Dr. Snow's request, the pump was disabled by removing its handle, which immediately ended the neighbourhood epidemic. It was discovered later that the well belonging to the pump was contaminated with sewage, which caused the outbreak in the neighborhood.

For an example of how data visualization can reveal model misspecification, consider Anscombe's quartet Anscombe (1973) shown in Figure 2. The four scatter plots all have identical summary statistics (i.e., means, standard deviations, and Pearson correlation coefficient). Only by visually inspecting the panels does it become obvious that the bivariate relation is fundamentally different for each panel (see also Matejka and Fitzmaurice, 2017).

1.1.3 Current Status

Since William Playfair (1759–1823) invented the first statistical graphs –such as line graphs and bar charts (Playfair, 1786)– , 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 (e.g., principal component analysis, or prior and posterior predictive checks), or for handling big data sets (e.g., through cluster analysis; Everitt et al., 2011). Technology now allows us to go beyond static visualizations and display the dynamic aspects of the data, for instance, by using the software packages R Shiny (Chang et al., 2020) or iNZight (iNZight Team, 2020).

1.1.4 Limitations

Despite the obvious benefits, data visualization also offers the opportunity to mislead, for instance, when displaying spurious patterns by either expanding the scale to minimize variation, or by minimizing the scale to accentuate differences (e.g., Cairo, 2019; Gelman, 2011; Wainer, 1984).

Figure 1: Recreation of Dr. Snow's map of the distribution of deaths from cholera. In this map, the points represent the homes of the deceased and the crosses represent the water pumps in the neighborhood. The contaminated water pump that triggered the cholera epidemic in the neighborhood is located on Broad Street. Reprinted with permission from *Pioneer maps of health and disease in England* (p. 174), by E. W. Gilbert, 1958, The Royal Geographical Society (with the Institute of British Geographers).

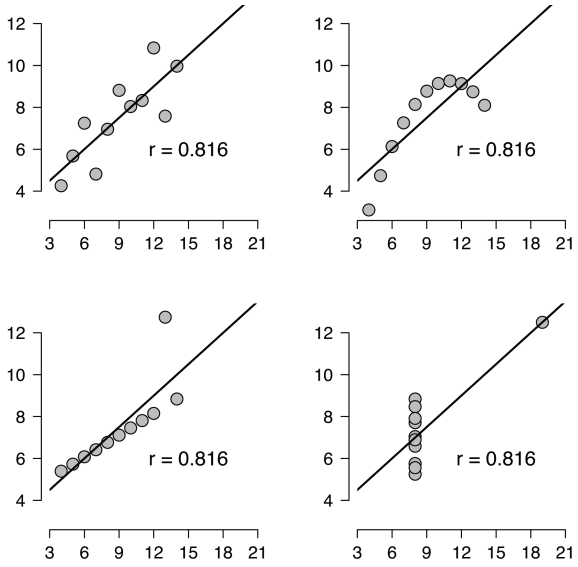

Figure 2: Anscombe's quartet emphasizes the importance of data visualization to detect model misspecification. Although the four data sets are equivalent in terms of their summary statistics, the Pearson correlation is only valid for the data set in the upper left panel.

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.

1.1.5 Guidelines

There are no uniform guidelines as to when and which graphical representations should be used. There is, however, a fundamental principle of good statistical graphics due to Tufte (1973, p.92): "Above all else show the data" (i.e., 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 (2018). 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" (p. 3).

Beyond that, guidelines depend on the individual aspects of the data (e.g., complexity of the data and experimental design) and context (cf. Diamond and Lerch, 1992); here we refer the interested reader to the numerous manuals describing good practices in graphical representation of statistical information (e.g., Chen et al., 2008; Cleveland and McGill, 1984; Gelman et al., 2002; Mazza, 2009; Tufte, 1973; Wilke, 2019; Wilkinson, 1999).

1.2 QUANTIFYING INFERENTIAL UNCERTAINTY

> "There is no excuse whatever for omitting to give a properly determined standard error. $(\cdots)$ All statisticians will agree with me here $(\cdots)$"
>
> Harold Jeffreys (1961, p. 410)

1.2.1 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.

1.2.2 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. (1988) 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 10-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).

1.2.3 Current Status

In virtually all statistics courses, students are taught to provide not only the summary of statistical tests (such as F-, t-, p-values and associated degrees of freedom), but also parameter point-estimates (e.g., regression weights, effect sizes) and their associated uncertainty

(e.g., 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 (Weissgerber et al., 2015), the social sciences (Hoekstra et al., 2006), and medicine (Cooper et al., 2002; Schriger et al., 2006) revealed that error bars, standard errors, or confidence intervals were not always presented. Also, popular metrics such as Cronbach's alpha (a measure of test score reliability) are virtually never presented with a measure of inferential uncertainty.

1.2.4 Limitations

We agree with Jeffreys's comment in the epigraph that there are no acceptable excuses for omitting a measure of inferential uncertainty in any report.

Although not a limitation per se, 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.

1.2.5 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 instance, the publication manual of the American Psychological Association (6th ed.) states: "When point estimates (e.g., sample means or regression coefficients) are provided, always include an associated measure of variability (precision), with an indication of the specific measure used (e.g., the standard error)," (p. 34). Also, the International Committee of Medical Journal Editors (2019) explicitly recommend to "[w]hen possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as

confidence intervals)" (p. 17).

1.3 ASSESSING DATA PREPROCESSING CHOICES

> "(· · ·) raw data do not uniquely give rise to a single data set for analysis but rather to multiple alternatively processed data sets, depending on the specific combination of choices—a many worlds or multiverse of data sets."
>
> Steegen et al. (2016, p. 702)

1.3.1 Description

By assessing the impact of plausible alternative data pre-processing choices (i.e., examining the "data multiverse", Steegen et al., 2016), the analyst determines the extent to which the finding under scrutiny is either fragile or sturdy.

1.3.2 Benefits and Example

A "data multiverse" analysis reveals the fragility or sturdiness of the finding under plausible alternative data pre-processing choices. This prevents researchers from falling prey to hindsight bias and motivated reasoning, which may lead them to unwittingly report only the pre-processing pipeline that yields the most compelling result (e.g., De Groot, 1956/2014; Simmons et al., 2011). 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. (2016) reexamined the results of Durante et al. (2013), who reported an interaction between relationship status (i.e., single or not) and menstrual cycle (i.e., fertile or not) on reported religiosity. After applying a series of 180 different data pre-processing procedures (e.g., 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.

1.3.3 Current Status

The idea of assessing sensitivity to data-preprocessing choices dates back at least to De Groot (1956/2014, p. 190) and Leamer (1985, p. 308) and was revived by Simmons et al. (2011) and by Steegen et al. (2016). In the field of functional magnetic resonance imaging, both Carp (2012) and Poldrack et al. (2017) emphasized the hidden influence of different plausible pre-processing pipelines. In psychology, recent applications are Bastiaansen et al. (2020) and Wessel et al. (2020). Nevertheless, the overwhelming majority of empirical articles does not report the results of a data multiverse analysis.

1.3.4 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 pre-processing choices and whether all pre-processing 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 (e.g., specification curves; Simonsohn et al., 2020).

1.3.5 Guidelines

Some specific guidelines on assessing data pre-processing choices are offered by Simmons et al. (2011, see 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 judgment about which options can be considered reasonable and will typically depend on the experimental design, the research question, and the researchers performing the research" (Steegen et al., 2016, p. 709). More general guidelines that relate exclusively to the reporting of pre-processing choices are given in the ASA Ethical Guidelines (2018). These mention that to insure 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" (p. 2).

1.4 REPORTING MULTIPLE MODELS

> "[A]ny approach that selects a single model and then makes inference conditionally on that model ignores the uncertainty involved in model selection, which can be a big part of overall uncertainty about quantities of interest."
>
> Kass and Raftery (1995, p. 26)

1.4.1 Description

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

1.4.2 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 modeling choices. Modeling 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. In addition, 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. (2015) 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. Consequently, they computed p-values for $2^{13} = 8,192$ models and examined the instability of the inference, which they call the "vibration of effects".

1.4.3 Current Status

Although the idea of the model multiverse dates back at least to De Groot (1956/2014) and Leamer (1985), most empirical researchers still base their conclusion on only a single analysis (but see Athey and Imbens, 2019; Levine and Renelt, 1992).

1.4.4 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 (i.e., Silberzahn

et al., 2018; Botvinik–Nezer et al., 2020). Also, multiverses 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). In addition, Del Giudice et al. (in press, p. 5) 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 and sheer size, trap whatever information is present in impenetrable displays and summaries."

1.4.5 Guidelines

Because the construction of the model multiverse depends on the knowledge and expertise of the analyst, it is challenging to provide general guidelines. For relatively simple regression models, however, clear guidelines do exist (e.g., Hoeting et al., 1999; Patel et al., 2015). Furthermore, Simonsohn et al. (2020) suggested a specification curve analysis, and Dragicevic et al. (2019) suggest interactive ways of presenting the results. The ASA Ethical Guidelines (2018) 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" (p. 3). The ASA, however, 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.

1.5 INVOLVING MULTIPLE ANALYSTS

> "The best defense against subjectivity in science is to expose it."
>
> Silberzahn et al. (2018, p. 354)

1.5.1 Description

By having multiple analysts independently analyze the same data set, the researcher can decrease the impact of analyst-specific choices regarding data pre-processing and statistical modeling.

1.5.2 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. (2018) where 29 teams of analysts examined, using the same dataset, whether the skin tone of soccer players influences their probability of getting a red card. While 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 pre-processing and statistical modeling (e.g., included covariates, link functions, assumption of hierarchical structure).

1.5.3 Current Status

A precursor to the multiple-analysts approach concerns the 1857 "Cuneiform competition", where four scholars independently translated a previously unseen ancient Assyrian inscription (Rawlinson et al., 1857). The overlap between their translations –sent to the Royal Asian Society in sealed envelopes, and 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 (Bastiaansen et al., 2020; Boehm et al., 2016; Botvinik–Nezer et al., 2020; Dutilh et al., 2019; Silberzahn et al., 2018; van Dongen et al., 2019).

1.5.4 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 data set, 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. While differences in opinion should be respected, there need to be ways to filter out analysis approaches that involve clear mistakes. An additional limitation concerns possible homogeneity of the analysts. For instance, 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. This potential limitations can be mitigated by selecting a diverse group of analysts and incorporating feedback and revision options in the process (Silberzahn et al., 2018), a round-table discussion (van Dongen et al., 2019) or, more systematically, a Delphi approach (Thangaratinam and Redman, 2005).

1.5.5 Guidelines

There are no explicit guidelines concerning the multiple-analysts approach. We propose that the optimal number of analysts to include depends on factors such as the complexity of the data, the importance of the research question (e.g., a clinical trial on the effectiveness of a new drug against COVID-19 warrants a relatively large number of analysts), and the probability that the analysts could reasonably reach a different conclusion (e.g., 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 (2018) 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 ($\cdots$) can arrive at different and potentially diverging judgments about the plausibility of different explanations" (p. 5).

1.6 INTERPRETING RESULTS MODESTLY

> "(· · ·) it is imperative in science to doubt; it is absolutely necessary, for progress in science, to have uncertainty as a fundamental part of your inner nature. To make progress in understanding, we must remain modest and allow that we do not know. Nothing is certain or proved beyond all doubt. "
>
> Richard Feynman (1956, p. 21)

1.6.1 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.

1.6.2 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. Also, 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" (Chadwick, 1932).

1.6.3 Current Status

Tukey (1962) already remarked that "Laying aside unethical practices, one of the most dangerous [$(\cdots)$ practices of data analysis $(\cdots)$] is the use of formal data-analytical procedures for sanctification, for the preservation of conclusions from all criticism, for the granting of an imprimatur." (p. 13). Almost 60 years later, an editorial in *Nature Human Behaviour* warns its readers about "conclusive narratives that leave no room for ambiguity or for conflicting or inconclusive results" (NHB Editorial, 2020, p. 1). Similarly, Simons et al. (2017) 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 of all, the frequency of stronger language (words like "amazing", "groundbreaking", "unprecedented") seemed to have increased in the last few decades (Vinkers et al., 2015). Secondly, dichotomization of findings (i.e., ignoring the uncertainty inherent to statistical inference) is common practice (e.g., Hoekstra et al., 2006; also see paragraph 4.3). Thirdly, textbooks (which are typically a reflection of current practice) on how to write papers often explicitly encourage authors to overclaim (e.g., Bem, 1987, van Doorn et al., in press).

1.6.4 Limitations

Publications and grants are important for scientific survival. Coupled with the fact that journals and funders often prefer groundbreaking 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 behavior (Yarkoni, 2018).

1.6.5 Guidelines

There are several ways we can contribute to increasing intellectual modesty. First of all, we could encourage intellectual modesty in others' work when we act as reviewers of papers and grant proposals (Hoekstra and Vazire, 2020). Since 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 (2020) 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 (p. 16) includes "Titles should not state or imply stronger claims than are justified (e.g., causal claims without strong evidence)".

Also, the ASA Guidelines (2018) 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" (p. 2).

1.7 SHARING DATA AND CODE

> "these and all your communications will be useless to me unless you can propose some practicable way or other of supplying me with Observations $(\cdots)$ I want not your calculations, but your Observations only."
>
> Sir Isaac Newton, 1695 (Kollerstrom and Yallop, 1995, p. 237)

1.7.1 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.

1.7.2 Benefits and Example

Since there are many different ways of processing and analyzing data (Silberzahn et al., 2018; Steegen et al., 2016), sharing code promotes reproducibility and encourages sensitivity analyses. Sharing data and code also allows 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 (1904) shared his data as an appendix to the article. A century later, this act of foresight enabled scientists to use this data set for both research and education. Because Spearman made his data publicly available, other researchers could establish the reproducibility and generalizability of the findings.

1.7.3 Current Status

Data sharing has never been easier. Public repositories offer free storage space for research materials, data (e.g., the Open Science Framework), and code (e.g., Github). While data sharing is not yet a general practice in most scientific fields, several recent initiatives (e.g., Open Data/Code/Materials badges, Kidwell et al., 2016), standards (TOP Guidelines, Nosek et al., 2015), journals (e.g., Scientific Data) and checklists (e.g., Transparency Checklist, Aczel et al., 2020b) are helping to promote this research practice. When sharing raw data is unfeasible, researchers can make aggregated data summaries available, for ex-

ample, the data used to generate certain plots or covariance matrices of involved variables.

1.7.4 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 (e.g., 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. It should be noted that these limitations should not apply to the analysis code as long as code is solely reflective of the researcher's analysis actions and is free of any data privacy issues.

1.7.5 Guidelines

An important principle of sharing data is that they should be Findable, Accessible, Interoperable, and Reusable (FAIR, Wilkinson et al., 2016). Several guides are available discussing the practical (e.g., Klein et al., 2018) and ethical (e.g., Alter and Gonzalez, 2018) aspects of data sharing. Researchers should follow the data sharing procedures and requirements of their fields (e.g., Wagenmakers et al., 2020; Taichman et al., 2017) and indicate the accessibility of the data in the research report (Aalbersberg et al., 2018; Nosek et al., 2015). The ASA Ethical Guidelines for Statistical Practice (2018) 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." (p. 5).

2 CONCLUDING COMMENTS

If the statistical literature is any guide, one may conclude that statisticians rarely agree with one another. For instance, the 2019 special issue in *The American Statistician* featured 43 articles on p-values, and in their editorial Wasserstein et al. (2019) 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 there is nevertheless much common ground among statisticians, specifically with respect to the ethical aspects of their profession. To explore 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 (e.g., 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" (Bennett, 1990, p. 162).

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. Also, instead of counting on them to be absorbed through osmosis, we believe 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.

AUTHOR CONTRIBUTIONS

Conceptualization: E.J. Wagenmakers, A. Sarafoglou, and B. Aczel.

Project Administration: B. Aczel.

Writing - Original Draft Preparation: E.J. Wagenmakers, A. Sarafoglou, C. Albers, J. Algermissen, S. Bahník, N. van Dongen, R. Hoekstra, D. Moreau, D. van Ravenzwaaij, A. Sluga, J. Tendeiro, and B. Aczel.

Writing - Review & Editing: E.J. Wagenmakers, A. Sarafoglou, S. Aarts, C. Albers, J. Algermissen, S. Bahník, N. van Dongen, R. Hoekstra, D. Moreau, D. van Ravenzwaaij, A. Sluga, F. Stanke, J. Tendeiro, and B. Aczel.

ACKNOWLEDGEMENTS

We are grateful to Nicole Lazar for her comments on a draft version. We also 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. Wagenmakers (283876), a Netherlands Organisation for Scientific Research (NWO) grant to A. Sarafoglou (406-17-568), as well as a Dutch

scientific organization Vidi grant from the NWO to D. van Ravenzwaaij (016.Vidi.188.001).

CONFLICTS OF INTEREST

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

References

Aalbersberg, I. J., Appleyard, T., Brookhart, S., Carpenter, T., Clarke, M., Curry, S., Dahl, J., DeHaven, A., Eich, E., Franko, M., Freedman, L., Graf, C., Grant, S., Hanson, B., Joseph, H., Kiermer, V., Kramer, B., Kraut, A., Karn, R. K., Lee, C., MacFarlane, A., Martone, M., Mayo-Wilson, E., McNutt, M., McPhail, M., Mellor, D., Moher, D., Mudditt, A., Nosek, B., Orland, B., Parker, T., Parsons, M., Patterson, M., Santos, S., Shore, C., Simons, D., Spellman, B., Spies, J., Spitzer, M., Stodden, V., Swaminathan, S., Sweet, D., Tsui, A., and Vazire, S. (2018). Making science transparent by default; Introducing the TOP statement. OSF Preprints.

Aczel, B., Hoekstra, R., Gelman, A., Wagenmakers, E., Klugkist, I. G., Rouder, J. N., Vandekerckhove, J., Lee, M. D., Morey, R. D., Vanpaemel, W., Dienes, Z., and van Ravenzwaaij, D. (2020a). Discussion points for Bayesian inference. *Nature Human Behaviour*, 4:561–566.

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, v., Benjamin, D., Chambers, C. D., Fisher, A., Gelman, A., Gernsbacher, M. A., Ioannidis, J. P., Johnson, E., Jonas, K., Kousta, S., Lilienfeld, S. O., Lindsay, D. S., Morey, C. C., Munafò, M., Newell, B. R., Pashler, H., Shanks, D. R., Simons, D. J., Wicherts, J. M., Albarracin, D., Anderson, N. D., Antonakis, J., Arkes, H., Back, M. D., Banks, G. C., Beevers, C., Bennett, A. A., Bleidorn, W., Boyer, T. W., Cacciari, C., Carter, A. S., Cesario, J., Clifton, C., Conroy, R. M., Cortese, M., Cosci, F., Cowan, N., Crawford, J., Crone, E. A., Curtin, J., Engle, R., Farrell, S., Fearon, P., Fichman, M., Frankenhuis, W., Freund, A. M., Gaskell, M. G., Giner-Sorolla, R., Green, D. P., Greene, R. L., Harlow, L. L., Hoces de la Guardia, F., Isaacowitz, D., Kolodner, J., Lieberman, D., Logan, G. D., Mendes, W. B., Moersdorf,

L., Nyhan, B., Pollack, J., Sullivan, C., Vazire, S., and Wagenmakers, E.-J. (2020b). A consensus-based transparency checklist. *Nature Human Behaviour*, 4:4–6.

Alter, G. and Gonzalez, R. (2018). Responsible practices for data sharing. *American Psychologist*, 73:146–156.

Amrhein, V., Greenland, S., and McShane, B. B. (2019). Retire statistical significance. *Nature*, 567:305–307.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27:17–21.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

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., et al. (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.

Bem, D. J. (1987). Writing the empirical journal. In Zanna, M. R. and Darley, J. M., editors, *The compleat academic: A practical guide for the beginning social scientist*, pages 171–201. Lawrence Erlbaum Associates, Mahwah, NJ.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A., Hadfield, J. D., Hedges, L. V., Held, L., Ho,

T.-H., Hoijtink, H., Jones, J. H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D., Morgan, S. L., Munafò, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2:6–10.

Bennett, J. H., editor (1990). *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*. Clarendon Press, Oxford.

Boehm, U., Hawkins, G. E., Brown, S. D., van Rijn, H., and Wagenmakers, E.-J. (2016). Of monkeys and men: Impatience in perceptual decision–making. *Psychonomic Bulletin & Review*, 23:738–749.

Botvinik–Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, A., Avesani, P., Baczkowski, B., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., Benoit, R., Berkers, R., Bhanji, J., Biswal, B., Bobadilla–Suarez, S., Bortolini, T., Bottenhorn, K., Bowring, A., Braem, S., Brooks, H., Brudner, E., Calderon, C., Camilleri, J., Castrellon, J., Cecchetti, L., Cieslik, E., Cole, Z., Collignon, O., Cox, R., Cunningham, W., Czoschke, S., Dadi, K., Davis, C., De Luca, A., Delgado, M., Demetriou, L., Dennison, J., Di, X., Dickie, E., Dobryakova, E., Donnat, C., Dukart, J., Duncan, N. W., Durnez, J., Eed, A., Eickhoff, S., Erhart, A., Fontanesi, L., Fricke, G. M., Galvan, A., Gau, R., Genon, S., Glatard, T., Glerean, E., Goeman, J., Golowin, S., González–García, C., Gorgolewski, K., Grady, C., Green, M., Guassi Moreira, J., Guest, O., Hakimi, S., Hamilton, J. P., Hancock, R., Handjaras, G., Harry, B., Hawco, C., Herholz, P., Her-

man, G., Heunis, S., Hoffstaedter, F., Hogeveen, J., Holmes, S., Hu, C.-P., Huettel, S., Hughes, M., Iacovella, V., Iordan, A., Isager, P., Isik, A. I., Jahn, A., Johnson, M., Johnstone, T., Joseph, M., Juliano, A., Kable, J., Kassinopoulos, M., Koba, C., Kong, X., Koscik, T., Kucukboyaci, N. E., Kuhl, B., Kupek, S., Laird, A., Lamm, C., Langner, R., Lauharatanahirun, N., Lee, H., Lee, S., Leemans, A., Leo, A., Lesage, E., Li, F., Li, M., Lim, P. C., Lintz, E., Liphardt, S., Losecaat Vermeer, A., Love, B., Mack, M., Malpica, N., Marins, T., Maumet, C., McDonald, K., McGuire, J., Melero, H., Méndez Leal, A., Meyer, B., Meyer, K., Mihai, P., Mitsis, G., Moll, J., Nielson, D., Nilsonne, G., Notter, M., Olivetti, E., Onicas, A., Papale, P., Patil, K., Peelle, J. E., Pérez, A., Pischedda, D., Poline, J., Prystauka, Y., Ray, S., Reuter–Lorenz, P., Reynolds, R., Ricciardi, E., Rieck, J., Rodriguez–Thompson, A., Romyn, A., Salo, T., Samanez–Larkin, G., Sanz–Morales, E., Schlichting, M., Schultz, D., Shen, Q., Sheridan, M., Shiguang, F., Silvers, J., Skagerlund, K., Smith, A., Smith, D., Sokol–Hessner, P., Steinkamp, S., Tashjian, S., Thirion, B., Thorp, J., Tinghög, G., Tisdall, L., Tompson, S., Toro–Serey, C., Torre, J., Tozzi, L., Truong, V., Turella, L., van't Veer, A. E., Verguts, T., Vettel, J., Vijayarajah, S., Vo, K., Wall, M., Weeda, W. D., Weis, S., White, D., Wisniewski, D., Xifra–Porxas, A., Yearling, E., Yoon, S., Yuan, R., Yuen, K., Zhang, L., Zhang, X., Zosky, J., Nichols, T. E., Poldrack, R. A., and Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582:84–88.

Brillinger, D. R. (2014). $\cdots$ how wonderful the field of statistics is $\cdots$. In Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J.-L., editors, *Past, Present, and Future of Statistical Science*, pages 65–I 72. CRC Press, Boca-Raton, FL.

Cairo, A. (2019). *How Charts Lie: Getting Smarter about Visual Information*. WW Norton & Company, New York.

Carp, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, 6:1–13.

Chadwick, J. (1932). Possible existence of a neutron. *Nature*, 129:312.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2020). shiny: Web application framework for R [Computer software]. `http://CRAN.R-project.org/package=shiny`.

Chen, C., Härdle, W., and Unwin, A., editors (2008). *Handbook of Data Visualization*. Springer Verlag, Berlin.

Cleveland, W. S. and McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79:531–554.

Cooper, R. J., Schriger, D. L., and Close, R. J. (2002). Graphical literacy: The quality of graphs in a large-circulation journal. *Annals of Emergency Medicine*, 40:317–322.

De Groot, A. D. (1956/2014). 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.

Del Giudice, M., Gangestad, S. W., and Steven, W. (in press). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*.

Diamond, L. and Lerch, F. J. (1992). Fading frames: Data presentation and framing effects. *Decision Sciences*, 23:1050–1071.

Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., and Chevalier, F. (2019). Increasing the transparency of research papers with explorable multiverse analyses. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Durante, K. M., Rae, A., and Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science*, 24:1007–1016.

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Krypotos, A.-M., et al. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review*, 26:1051–1069.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. John Wiley & Sons, Chichester.

Feynman, R. P. (1956). The relation of science and religion. *Engineering and Science*, 19:20–23.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182:389–402.

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, 13:755–779.

Gelman, A. (2011). Why tables are really much better than graphs. *Journal of Computational and Graphical Statistics*, 20:3–7.

Gelman, A., Pasarica, C., and Dodhia, R. (2002). Let's practice what we preach: Turning tables into graphs. *The American Statistician*, 56:121–130.

Gilbert, E. W. (1958). Pioneer maps of health and disease in England. *The Geographical Journal*, 124:172–183.

Good, I. J. (1971). 46656 varieties of Bayesians. *The American Statistician*, 25:62–63.

Harlow, L. L., Mulaik, S. A., and Steiger, J. H., editors (1997). *What if There Were No Significance Tests?* Lawrence Erlbaum, Mahwah (NJ).

Healy, K. and Moody, J. (2014). Data visualization in sociology. *Annual Review of Sociology*, 40:105–128.

Heathcote, A., Brown, S. D., and Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In *An Introduction to Model-Based Cognitive Neuroscience*, pages 25–48. Springer Verlag.

Hoekstra, R., Finch, S., Kiers, H. A., and Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p-values. *Psychonomic Bulletin & Review*, 13:1033–1037.

Hoekstra, R. and Vazire, S. (2020). Intellectual humility is central to science: Some practices to aspire to. *PsyArXiv*.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, pages 382–401.

International Committee of Medical Journal Editors (2019). Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. `http://www.icmje.org/icmje-recommendations.pdf`.

iNZight Team (2020). iNZight (Version 4.0.2.) [Computer software]. `https://inzight.nz`.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford, UK, 3 edition.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kerman, J., Gelman, A., Zheng, T., and Ding, Y. (2008). Visualization in Bayesian data analysis. In Chen, C., Härdle, W., and Unwin, A., editors, *Handbook of Data Visualization*, pages 709–724. Springer Verlag, Berlin.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess–Holden, C., Errington, T. M., Fiedler, S., and Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low cost, effective method for increasing transparency. *PLOS Biology*, 14:e1002456.

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Hofelich Mohr, A., IJzerman, H., Nilsonne, G., Vanpaemel, W., and Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, 4:1–15.

Kollerstrom, N. and Yallop, B. D. (1995). Flamsteed's lunar data, 1692–95, Sent to Newton. *Journal for the History of Astronomy*, 26:237–246.

Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review*, 75:308–313.

Levine, R. and Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *The American Economic Review*, pages 942–963.

Matejka, J. and Fitzmaurice, G. (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1290–1294.

Mazza, R. (2009). *Introduction to information visualization*. Springer Science & Business Media, London.

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73:235–245.

Merton, R. K. (1973). The normative structure of science (1942). In Merton, R. K., editor, *The Sociology of Science: Theoretical and Empirical Investigations*, pages 267–278. University of Chicago Press, Chicago, IL.

NHB Editorial (2020). Tell it like it is. *Nature Human Behaviour*, 4:1.

Nosek, B., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T. A., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Levy Paluck, E., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E.-J., Wilson, R., and Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348:1422–1425.

Patel, C. J., Burford, B., and Ioannidis, J. P. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68:1046–1058.

Playfair, W. (1786). *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*. Corry, London. Re-published in Wainer, H. and Spence, I. (eds.), The Commercial and Political Atlas and Statistical Breviary, 2005, Cambridge University Press.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., and Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18:115–126.

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., Morgan, A. C., Pentland, A., Polimis, K., Raes, L., Rigobon, D. E., Roberts, C. V., Stanescu, D. M., Suhara, Y., Usmani, A., Wang, E. H., Adem, M., Alhajri, A., AlShebli, B., Amin, R., Amos, R. B., Argyle, L. P., Baer-Bositis, L., Büchi, M., Chung, B.-R., Eggert, W., Faletto, G., Fan, Z., Freese, J., Gadgil, T., Gagné, J., Gao, Y., Halpern-Manners, A., Hashim, S. P., Hausen, S., He, G., Higuera, K., Hogan, B., Horwitz, I. M., Hummel, L. M., Jain, N., Jin, K., Jurgens, D., Kaminski, P., Karapetyan, A., Kim, E. H., Leizman, B., Liu, N., Möser, M., Mack, A. E., Mahajan, M., Mandell, N., Marahrens, H., Mercado-Garcia, D., Mocz, V., Mueller-Gastell, K., Musse, A., Niu, Q., Nowak, W., Omidvar, H., Or, A., Ouyang, K., Pinto, K. M., Porter, E., Porter, K. E., Qian, C., Rauf, T., Sargsyan, A., Schaffner, T., Schnabel, L., Schonfeld, B., Sender, B., Tang, J. D., Tsurkov, E., van Loon, A., Varol, O., Wang, X., Wang, Z., Wang, J., Wang, F., Weissman, S., Whitaker, K., Wolters, M. K., Woon, W. L., Wu, J., Wu, C., Yang, K., Yin, J., Zhao, B., Zhu, C., Brooks-Gunn, J., Engelhardt, B. E., Hardt, M., Knox, D., Levy, K., Narayanan, A., Stewart, B. M., Watts, D. J., and McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117:8398–8403.

Schriger, D. L., Sinha, R., Schroter, S., Liu, P. Y., and Altman, D. G. (2006). 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. *Annals of Emergency Medicine*, 48:750–756.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, u., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., Fong, N., Gamez–Djokic, M., Glenz, A., Gordon–McKeon, S., Heaton, T. J., Hederos, K., Heene, M., Hofelich Mohr, A. J., F., H., Hui, K., Johannesson, M., Kalodimos, J., Kaszubowski, E., Kennedy, D. M., Lei, R., Lindsay, T. A., Liverani, S., Madan, C. R., Molden, D., Molleman, E., Morey, R. D., Mulder, L. B., Nijstad, B. R., Pope, N. G., Pope, B., Prenoveau, J. M., Rink, F., Robusto, E., Roderique, H., Sandberg, A., Schlüter, E., Schönbrodt, F. D., Sherman, M. F., Sommer, S. A., Sotak, K., Spain, S., C., S., Stafford, T., Stefanutti, L., Tauber, S., Ullrich, J., Vianello, M., Wagenmakers, E.-J., Witkowiak, M., Yoon, S., and Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1:337–356.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False–positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22:1359–1366.

Simons, D. J., Shoda, Y., and Lindsay, D. S. (2017). Constraints on generality (cog): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12:1123–1128.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2020). Specification curve analysis. *Nature Human Behaviour*, 4:1208–1214.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293.

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11:702–712.

Strack, F., Martin, L. L., and Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54:768–777.

Taichman, D. B., Sahni, P., Pinborg, A., Peiperl, L., Laine, C., James, A., Hong, S.-T., Haileamlak, A., Gollogly, L., Godlee, F., Frizelle, F. A., and Flor, F. (2017). Data sharing statements for clinical trials: A requirement of the International Committee of Medical Journal Editors. *JAMA*, 317:2491–2492.

Thangaratinam, S. and Redman, C. W. (2005). The delphi technique. *The Obstetrician & Gynaecologist*, 7:120–125.

Tufte, E. R. (1973). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33:1–67.

Tukey, J. W. (1977). *Explanatory Data Analysis*. Addison–Wesley, Reading, MA.

van Dongen, 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., and Wagenmakers, E.-J. (2019). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician*, 73:328–339.

van Doorn, J., van den Bergh, D., Dablander, F., van Dongen, N., Derks, K., Evans, N. J., Gronau, Q. F., Haaf, J. M., Kunisato, Y., Ly, A., Marsman, M., Sarafoglou, A., Stefan, A., and Wagenmakers, E. (in press). Strong public claims may not reflect researchers' private convictions. *Significance*.

Vinkers, C. H., Tijdink, J. K., and Otte, W. M. (2015). Use of positive and negative words in scientific pubmed abstracts between 1974 and 2014: Retrospective analysis. *BMJ*, 351:h6467.

Wagenmakers, E.-J., Kucharsky, S., and the JASP Team, editors (2020). *The JASP Data Library*. JASP Publishing, Amsterdam.

Wainer, H. (1984). How to display data badly. *The American Statistician*, 38:137–147.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p–values: Context, process, and purpose. *The American Statistician*, 70:129–133.

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$". *The American Statistician*, 73:1–19.

Weissgerber, T. L., Milic, N. M., Winham, S. J., and Garovic, V. D. (2015). Beyond bar and line graphs: Time for a new data presentation paradigm. *PLoS Biology*, 13:e1002128.

Wessel, I., Albers, C., Zandstra, A. R. E., and Heininga, V. E. (2020). A multiverse

analysis of early attempts to replicate memory suppression with the Think/No-think task. *Memory*, 28:870–887.

Wilke, C. O. (2019). *Fundamentals of data visualization: A primer on making informative and compelling figures*. O'Reilly Media, Sebastopol, CA.

Wilkinson, L. (1999). *The Grammar of Graphics*. Springer Science & Business Media, New York.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzales–Beltran, A., Gray, A. J. G., Groth, P., Goble, C., S., G. J., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca–Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3.

Yarkoni, T. (2018). No, it's not the incentives–it's you. `https://www.talyarkoni.org/blog/2018/10/02/no-its-not-the-incentives-its-you/`.

A SEVEN MERTONIAN STATISTICAL PROCEDURES

This appendix outlines how how each of the seven procedures discussed in the main manuscript fullfill the Mertonian norms. An overview of how each statistical practice adheres to Merton's four norms of science, are given in Table 1.

Table 1: How various statistical practices adhere to Merton's four norms of science.

	Communalism	Universalism	Disinterestedness	Organized Skepticism
1. Visualizing Data	✔		✔	✔
2. Quantifying Inferential Uncertainty	✔		✔	✔
3. Assessing Data Preprocessing Choices	✔		✔	✔
4. Reporting Multiple Models	✔		✔	✔
5. Involving Multiple Analysts		✔	✔	✔
6. Interpreting Results Modestly			✔	✔
7. Sharing Data and Code	✔	✔	✔	✔

A.1 Visualizing Data

Proper data visualization embodies the Mertonian norms of communalism, disinterestedness, and organized skepticism. Well-designed visualizations show at a glance the key aspects of the data. This makes them an important part of a complete and open communication, thus serving the goal of communalism. 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 find alternative ways of

interpreting the results and analyzing the data, thus serving the goals of disinterestedness and organized skepticism.

A.2 Quantifying Inferential Uncertainty

Quantifying inferential uncertainty serves the Mertonian norms of communalism, disinterestedness, and organized skepticism. Acknowledging inferential uncertainty (e.g., by presenting standard errors or confidence intervals) contributes to open and transparent communication. In addition, quantifying inferential uncertainty signals that researchers are openly acknowledging the extent to which their measurements are imprecise, especially when 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.

A.3 Assessing Data Preprocessing Choices

Reporting the sensitivity of the conclusions to changes in data pre-processing connects to the norms of communalism, disinterestedness and organized skepticism. When researchers share the results from only a single data pre-processing pipeline, they may unintentionally hide potentially important information. In addition, by reporting on the fragility of their finding, researchers underscore their disinterestedness, as a biased researcher may be tempted to apply only the pre-processing pipeline that yields the most flattering result. Finally, if a result proves sensitive to particular pre-processing choices, this warrants skepticism and may initiate a debate on the importance and plausibility of relevant data pre-processing choices. As mentioned by Leamer (1985, p. 308):

"Decentralized studies of fragility are common whenever an inference matters enough

to attract careful scrutiny. $(\cdots)$ These disorganized studies of fragility are inefficient, haphazard, and confusing. What we need instead are organized sensitivity analyses. We must insist that all empirical studies offer convincing evidence of inferential sturdiness. We need to be shown that minor changes in the list of variables do not alter fundamentally the conclusions, nor does a slight reweighting of observations, nor corrections for dependence among observations, etcetera, etcetera."

A.4 Reporting Multiple Models

Reporting the sensitivity of the conclusions to changes in statistical modeling connects to the norms of communalism, disinterestedness and organized skepticism. The rationale mirrors that outlined in the previous section, "Assessing data pre-processing choices".

A.5 Involving Multiple Analysts

The multiple-analysts approach embodies the Mertonian norms of universalism, disinterestedness, and organized skepticism. The approach can reveal whether different (teams of) analysts reach converging or diverging conclusions from the same data set. By including other analysts with different backgrounds and interests, the potential impact of self-interest of any single analyst is counteracted, thus serving the aim of disinterestedness. Finally, the multiple-analysts approach stimulates skepticism by bringing to light alternative statistical perspectives on the data.

A.6 Interpreting Results Modestly

A modest interpretation of the results is in line with the norms of disinterestedness and organized skepticism. Disinterested analysts arguably have little need to exaggerate claims,

impress reviewers, and downplay signs of model misfit. Analysts who facilitate organized skepticism do not attempt to suppress doubt –they are not defensive, and they do not wish to protect their work against well-intentioned scrutiny from their peers.

A.7 Sharing Data and Code

Open data should be a norm of science according to all Mertonian imperatives. All interested researchers should have access to relevant, properly anonymized data. All interested researchers should have access to relevant, properly anonymized data –those who collected the data should not have exclusive rights to the observations. As long as research is done for the advancement of science, the benefits of the public outweigh the interests of the scientist. All secrecy about the data is a limitation to knowledge accumulation and violates the ethos of science. The observations and not the authority of the researchers should be the basis of scientific claims. Importantly, sharing data allows skeptical eyes to scrutinize the results, promoting quality control.