# Response to discussion of "Beyond subjective and objective in statistics"*

Andrew Gelman† Christian Hennig‡

1 June 2017

As practicing scientists it is natural to feel that we are too busy for philosophy. We are grateful to have found so many thoughtful discussants who agree with us on the value of reflecting on the values, often left unstated, which underlie statistical theory, methods, and practice.

We wrote our article because we believe that words matter. From one direction, we feel that outmoded attitudes of objectivity and subjectivity have impeded the work of statisticians, by discouraging some good practices and by motivating some bad practices. At the same time, we recognize the real concerns underlying the decades-long debates regarding objective and subjective perspectives in statistics, and we believe that an uncovering of these more fundamental concerns, as represented by the list of virtues in Table 1 of our article, can help us do better by freeing ourselves from artificial restrictions and by suggesting new directions.

But there are precedents for the expansion of the philosophical frameworks of statistics. Consider exploratory data analysis, which was traditionally considered to be separate from academic statistical methods but which has been brought into the fold via the formalizations of Tukey (1977), Rubin (1984), Gelman (2003), Wilkinson (2005), and Wickham (2017). It's not just that we now have a language to talk about statistical graphics in the context of models; we also have many valuable tools which allow us to learn from, check, and improve our models in ways that were not accessible when graphics was taken merely as good practice without a theoretical structure. For a completely different example, the framework of missing data has been used to systematize causal inference (Rubin, 1974). More recently, the replication crisis in psychology has stimulated new thinking regarding the interplay between statistical inference and experimental design in the context of the scientific publication process (Simmons, Nelson, and Simonsohn, 2011, Button et al., 2012).

The common theme in all the above examples is the parallel development of critiques of existing practice, ideas for improvements, and philosophical or theoretical developments. We view this discussion as a step in this process: our paper was an attempt to jolt the ideas of statistical objectivity and subjectivity into the modern world, and the discussions are a necessary course correction. Indeed, this conversation exemplifies several of the virtues tabulated in our article, including full communication (virtue V1(c)), openness to criticism and exchange (virtue V3(c)), and awareness of multiple perspectives and context dependence (virtues V5 and V6). The largest challenge in such a discussion may be correspondence to observable reality (virtue V4). We spelled out connections to statistical practice in section 4 of our article, and we hope to keep these and other applied examples in our minds in this discussion and moving forward.

We will go through the 54 (!) discussions in what seems to us to be a logical order, recognizing that space constraints stop us from being able to respond to all the issues that arose in the discussion.

We are delighted that the discussion contains many valuable additional suggestions. The list of meta-statistical considerations provided by Dawid is particularly rich and deserves a paper on its own.

We begin with reactions to our general setup. Vanpaemel, Tuerlinckx, and De Boeck point out that statistical discussions are typically framed in the context of a single study or a single analysis,

*To appear in the *Journal of the Royal Statistical Society*.

†Department of Statistics and Department of Political Science, Columbia University, New York.

‡Department of Statistical Science, University College London.

but that statistics by its nature is concerned with ensembles. In any example, we should consider our design and analysis choices in the context of other problems that we might be studying. This is a frequentist idea to consider the properties of statistical methods relative to a reference set, and it is also Bayesian in that the elements of this set can be assigned probabilities, to form a prior distribution or hierarchical model representing a larger class of problems.

McConway and Cox, Thall, and Porcu, Alvarez, and Carrion, connect Dawid's meta-statistics to another meta-idea, that the principles of transparent, aware science are also relevant to scientific communication and publication. We agree with Longford that the reporting of inferential summaries such as point estimates, uncertainties, and hypothesis tests involve choices which should be made with awareness of applied goals and with understanding of who will read these summaries and how they should be used.

Also related is the point made by Grant and Firth, and by Harper-Donnelly and Donnelly, that transparency is great, but transparency plus explanation is even better. Any good statistical analysis comes with scientific explanation—and, again, the virtues discussed in our paper (which in turn are derived from introspection and reflection upon what we view to be good statistical practice) are relevant to the explanation of quantitative findings. Future progress on exploratory model analysis (Urbanek, 2006, Wickham, 2006) may benefit from a more careful elaboration of goals and trade-offs among our list of virtues, for example following up on Morey's remark that increasing selectivity may decrease sensitivity. From the other direction, Josse, alone and with colleagues, argues that a lack of clarity in explanation could be a natural response of researchers to the incentives of publications that favor the illusion of certainty and objectivity. As Suarez writes, one gets better at statistical judgement with practice, and it helps to work within a framework that allows for such experimentation. One motivation for writing our article was to engage the philosophy community in this interplay we see between the philosophy and practice of statistics.

Stigler and Stehlik argue that the terms "subjectivity" and "objectivity" actually do convey opposing approaches to learning from data and thus we have no need to discard these in favor of our longer list of virtues. From a purely intellectual perspective, we might agree; but as noted above, we feel that arguments surrounding these terms have polluted the discourse in statistics, and so we prefer to focus on what we consider to be more fundamental goals. Our list of virtues in many settings is aspirational rather than descriptive; then again, "objectivity" is aspirational too, and we believe our aspirations are more understandable, more realistically attainable, and more useful in grappling with the hard problems of statistics. We do, however, agree with Stone that "honesty and transparency are not enough" (Gelman, 2017) and we did not mean to imply that complete adherence to our list of virtues—even if this were logically possible—would resolve all or even most statistical problems. There will always remain many challenges in computing, mathematical understanding of statistical methods, communication, measurement, and many other statistical problems. What we hope to get from our framework is a clearer view to help us to avoid various roadblocks that have arisen from slavish following of outdated philosophical principles.

Due to lack of space, some issues received less attention from us than they deserve. Major examples are nonparametric statistics, machine learning, and the M-open perspective in Bayesian statistics (as highlighted by Robert, Marin et al., and Meila). Our intention was certainly not to dismiss these approaches, which are attractive in many applications, particularly where prediction is a major aim.

Model assumptions are often presented in statistics in a misleading way, as if the application of methods would require assumptions to be true. But reality will typically differ from formal model assumptions (see also Hennig, 2010). Models set up idealized situations in which methods can be shown to work well. In this way, models contribute to transparency and understanding. But model-

based methodology can well be applied to data that are not really random (as raised by Robert). Model checks cannot make sure that the models are "true"; they help to avoid inappropriate and misleading analysis of the data. Romeijn's idea that true models will lead to true results, viewing statistics as logic, runs into problems here.

Methods that are not based on probability models perform well in some important tasks, but they come with their own implicit assumptions. We would argue that such methods work because they allow the flexible use of masses of data, which is often possible because the models are fit using regularization rather then pure optimization. Regularization in turn requires assumptions (or choices), which should be transparent and use subject-matter knowledge where possible, advice which we think is consistent with the experience of Murtagh on stable methods for big-data analytics.

One issue with supposedly assumption-free methodology is that it can discourage researchers from clarifying their specific research aims. In cluster analysis, for example, different research aims require different definitions of clusters, whereas model- and tuning-free methods are often advertised as making such decisions without "subjective" researcher input; see Hennig (2015). That said, we welcome assumption-light or model-free theoretical guarantees as mentioned by Meila. Lund and Iyer sketch a framework to analyze the interplay between data, assumptions and interpretation.

A number of contributions challenge our attitude toward realism and truth (Marin et al., Paul, Romeijn), sometimes together with questioning consensus or stability as virtues (Robert, O'Rourke, Thall, VanderWeele). Realism and the meaning of truth have been controversial issues in philosophy for ages, and we will not solve these controversies. "How things really are," as O'Rourke cites Peirce, is ultimately inaccessible to human beings. We can do experiments, make observations, and communicate them. The idea of some kind of objective, observer-independent reality ultimately relies on agreement about observations, in other words, some kind of consensus. Communication and consensus (and also stability and reproducibility; see Hannig's discussion) are crucial for making statements about reality that are claimed to hold for general observers. They are even crucial to set up the conditions for such statements such as agreed measurement procedures. "Truth" can certainly be meaningfully used within systems of communication in which there is agreement about how to establish it, for example within mathematics and logic or referring to measurements. The idea of truth applied to our models and many of our theories, though, leads us outside this domain. Betancourt and VanderWeele each take us on in a slightly different way, arguing that consensus should not be considered a virtue in itself but rather is—where it is appropriate at all—a consequence of other virtues. On the other hand, Thall and O'Rourke remind us that a consensus is not so valuable it is obtained too easily by following potentially misguided conventions. We still think of consensus as a valuable goal even it can be abused in group decision making. Consensus does not always exist, but people are rightly troubled by nagging disagreements. Jukola's addition that institutional and social conditions of inquiry should be taken into account is valuable. Boulesteix and Strasser deal with the tension between multiple perspectives, stability and consensus, using ensemble approaches to bring multiple perspectives together, but also highlighting the danger that individual perspectives should not be driven by partiality.

Robert and Bandyopadhyay are concerned with the logical holes inherent in the "falsificationist Bayesian" perspective. We're concerned about this too! From Gelman (2011); "My point here is not to say that my preferred methods are better than others but rather to couple my admission of philosophical incoherence with a reminder that there is no available coherent alternative." We agree with Sprenger that it is only rarely that a Bayesian prior distribution completely mirrors a statistician's or researcher's beliefs; rather, priors and data-generation models are imperfect codings of some subset of available information and thus are inherently subject to revision, and we agree

with Morey that the practice or principle parsimony fits awkwardly within our framework.

Moores points out that, following Lakatos rather than Popper, we are typically interested in improving rather than falsifying our models, which raises the question of what procedures should be used in the improvement step, considering that on Cantorian grounds we have already ruled out the existence of a super-model that would allow model building, inference, and improvement to all be done in one super-Bayesian step. Draper's approach using explicitly conditioning is, we hope, a step in making our practices more logically and statistically coherent.

Maclaren asks whether we think of robustness in the "Boxian" sense as a property of a good model that is flexible and not brittle, or in the "Tukeyian" sense as a set of operations to be performed on data. We think the two views are complementary, in that Boxian robust procedures can be considered as models that instantiate Tukeyian data-trimming procedures—and vice versa.

Wynn argues that Bayesians should stop worrying about objectivity and instead should embrace context dependence and multiple perspectives (virtues V5 and V6) by considering the iterative process of model building, checking, and improvement as occurring not inside the statistician's head but rather within a larger community of scientists and interested parties. We agree. But Celeux also has a good point when he says that the complexity of hierarchical Bayesian modeling makes transparency more difficult. Section 4.1 of our paper demonstrates how the specification of a prior distribution can be made more transparent by open admission of the process by which the prior has been constructed, but we agree that more work should be done in this area to routinize that sort of transparent exposition. Beyond this, as French notes, the most important aspects of a model are often "structural" and encode substantive models or assumptions; we should have a way of explicating their sources too.

From a different Bayesian direction, Jewson argues that, in the real-world "M-open" scenario in which the true data generating process is not included in the class of models used to fit the data, it could be possible to improve upon straight Bayesian updating, thus introducing another element of researcher's choice into data analysis. We suppose this could be incorporated into the usual Bayesian framework by instituting a loss function that captures which aspects of future data are of most interest. Relatedly, Leonelli and Smith connect the idea of "institutional decision analysis" (see section 3.1 of our article) to their work on decision support systems, which represent another way to transparently incorporate additional information and perspectives into decision making.

Mateu goes even further, not just recommending context awareness but disagreeing with our characterization of statistics as a "science of defaults" (Gelman, 2014a). In response we, along with Vukcevic, Moreno-Betancur, and Carlin, point to the world of practitioners in data collection and data analysis, many with little formal statistical training, who use available methods. Those of us who write software know that default settings are important, and much of our own research involves the development of methods that will work well the first time in a wide variety of realistic examples, and which are loud rather than quiet in their failure modes, so as to warn users away from some of the more disastrous consequences of inappropriate modeling and analysis choices. We appreciate that Wagenmakers likens our paper to Disneyland and we hope he will forgive us our "flirtation with frequentism" by recognizing the practical value for methodologists and textbook writers to get some sense of the statistical properties of our methods—that is, the consequences should our recommendations be followed in some specified distribution of cases.

Crane agrees with us that the terms "objective" and "subjective" have lost their value, but he would go one further than us by not attempting to categorize virtues at all. We agree with Crane that, ultimately, no principle is sacrosanct. For example one might think that correspondence with observed reality is essential to all science, but there can be value in purely theoretical work with no direct links to reality at all. Transparency is often beneficial but sometimes our methods outpace

our understanding, and there can be value in black-box methods that just seem to work. And so on. However, we do think our list of virtues is useful, not as a checklist or requirement for published work but as an elaboration and exposition of the real concerns underlying all that talk about subjectivity and objectivity in statistics.

Vandemeulebroecke asks how our list of virtues would apply in confirmatory settings in drug development where "emphasis is placed on type I error control and pre-specification of analyses." To the extent that researchers do care about such things, we would connect error control to impartiality (virtue V3) and pre-specification to open planning and following agreed protocols (virtue V1(b)). The point here is not to match every statistical method to a virtue, but to understand that methods are devised and used to satisfy some goals—in this case, a desire to avoid making poor decisions based on noisy data. However, given that in reality effects are not exactly zero, we should be able to satisfy this aim in the context of a more realistic model of nonzero, varying effects, which in turn will require some quantification of the costs and benefits of different decision options. Rather than setting a family-wide error threshold of 0.05 or 0.01, implicitly motivated by some appeal to consensus (virtue V2(b)), we recommend that users determine decision thresholds based on real-world contexts and aims (virtue V6(a)). This is an example of the larger point raised by Zyphur and Pierides, that our virtues can conflict with each other, but recognizing and adjudicating these conflicts can help us better understand the motivations for using different statistical rules.

King points out that a systematic study of goals and virtues could be appropriate not just for probability modeling (the focus of our article) but also when studying graphical communication and exploratory data analysis. Bartholomew and Kumar each ask about extensions in a different way, to more formally consider the relation between finite-population inference, the statistical theory of sampling, real-world surveys; our only thought here is that the relevant theory should address applied goals as well as the sampling process itself. And we agree with Vanpaemel, Tuerlinckx, and De Boeck that the same principles occur for measurement and design as with data analysis: rather than hiding one's decisions under the rug of convention or presumed objectivity, we believe it is better to be explicit about options and choices.

As Winkler explains, statisticians work at many levels: applied data collection and analysis, development of statistical theory and methods, construction of software, and, not least, communication and teaching. The relevance of the virtues we list for applied statistics may be somewhat different for those other tasks. Barker emphasizes that probability models are by their nature inexact, which motivates skepticism, a virtue that did not make it onto our list but which implicitly appears in many places, most notably openness to criticism (virtue V3(c)). Barker's concerns (i), (ii), and (iii) correspond to our virtues V1(c), V7(a), and V3(c).

Gepp, pointing to the literature on machine learning, discusses a way in which our virtues can interfere with each other, at least in the short term, so that direct pursuit of model improvement can reduce the correspondence of the fitted model to observed reality. In applied problems, there are many potential sources of information, and we think about the applied context when considering which variables to go to the trouble of measuring and modeling. Wijayatunga notes the importance of prediction accuracy as an criterion that should help with consensus—as long as there can be agreement on the outcomes to be predicted and the corpus of problems to average over. In a similar vein, Hannig explains how virtues such as transparency and stability can have different meanings under different inferential philosophies. Vukcevic, Moreno-Betancur, and Carlin emphasize that the steps of statistical analysis (as well as design and data collection) are guided by the applied goals of any project. Statistics textbook writers tend to take goals for granted and then jump right into the modeling and data analysis without always specifying the connection between assumptions and goals.

Commenting on two of the more specific statistical issues, the notion of outliers (branded as "nonsense in an M-open perspective" by Marin et al.) was used in our presentation to illustrate how background information such as regarding to what extent the measurements in question are prone to gross errors has implications on data analysis. We think that outliers are an important concept in the interpretation of data in many applications, regardless of what methodology and models are applied. Montanari asks about the connection between nonidentifiability and multiple perspectives. We think it important with nonidentified or weakly identified modes to acknowledge uncertainty, so that if one solution presented is is mathematically equivalent to many others, the results cannot be distinguished, and it would be inappropriate to pick one solution and to sell it as the unique answer. On the other hand, it is legitimate to present the result as compatible with multiple perspectives, and to select one for being, for example, the easiest possible interpretation.

Ultimately we agree with VanderWeele and Paul regarding the goal of a "consensus-based, rigorously vetted, scientific representation of the world." Much of the tension in our article, the discussants, and decades of earlier commentary on objectivity and subjectivity in statistics, comes from the goal of using flawed models and approximate methods to solve real problems and learn general truths. We hope that a better understanding of "the unreasonable effectiveness of statistics" (to paraphrase Wigner, 1960) will give us insights into how to make these tools even more effective.

## Additional references


Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 365-–376.

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review* **71**, 369–382.

Gelman, A. (2011). Induction and deduction in Bayesian data analysis. *Rationality, Markets and Morals*, special topic issue "Statistical Science and Philosophy of Science: Where Do (Should) They Meet In 2011 and Beyond?", ed. Deborah Mayo, Aris Spanos, and Kent Staley.

Gelman, A. (2017). Honesty and transparency are not enough. *Chance* **30** (1), 37–39.

Hennig, C. (2015) What are the true clusters? *Pattern Recognition Letters* **64**, 53–62.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

Urbanek, S. (2006). *Exploratory Model Analysis: An Interactive Graphical Framework for Model Comparison and Selection*. Books on Demand Gmbh.

Wickham, H. (2006). Exploratory model analysis with R and GGobi. Technical report. `http://had.co.nz/model-vis/2007-jsm.pdf`.

Wickham, H. (2017). The tidyverse. `http://tidyverse.org`.

Wilkinson, L. (2005). *The Grammar of Graphics*, second edition. New York: Springer-Verlag.

Wigner, E. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Communications in Pure and Applied Mathematics* **13**, 1–13.