American Journal of Epidemiology

Vol. 186, No. 6
DOI: 10.1093/aje/kwx259
Advance Access publication:
August 21, 2017

Invited Commentary

Invited Commentary: The Need for Cognitive Science in Methodology

Sander Greenland*

* Correspondence to Dr. Sander Greenland, Department of Epidemiology, School of Public Health, University of California, Los Angeles, Los Angeles, CA 90095 (e-mail contact only; e-mail: lesdomes@ucla.edu).

Initially submitted June 6, 2017; accepted for publication June 7, 2017.

There is no complete solution for the problem of abuse of statistics, but methodological training needs to cover cognitive biases and other psychosocial factors affecting inferences. The present paper discusses 3 common cognitive distortions: 1) *dichotomania*, the compulsion to perceive quantities as dichotomous even when dichotomization is unnecessary and misleading, as in inferences based on whether a *P* value is "statistically significant"; 2) *nullism*, the tendency to privilege the hypothesis of no difference or no effect when there is no scientific basis for doing so, as when testing only the null hypothesis; and 3) *statistical reification*, treating hypothetical data distributions and statistical models as if they reflect known physical laws rather than speculative assumptions for thought experiments. As commonly misused, null-hypothesis significance testing combines these cognitive problems to produce highly distorted interpretation and reporting of study results. Interval estimation has so far proven to be an inadequate solution because it involves dichotomization, an avenue for nullism. Sensitivity and bias analyses have been proposed to address reproducibility problems (*Am J Epidemiol.* 2017;186(6):646–647); these methods can indeed address reification, but they can also introduce new distortions via misleading specifications for bias parameters. *P* values can be reframed to lessen distortions by presenting them without reference to a cutoff, providing them for relevant alternatives to the null, and recognizing their dependence on all assumptions used in their computation; they nonetheless require rescaling for measuring evidence. I conclude that methodological development and training should go beyond coverage of mechanistic biases (e.g., confounding, selection bias, measurement error) to cover distortions of conclusions produced by statistical methods and psychosocial forces.

behavioral economics; bias analysis; cognitive bias; motivated reasoning; nullism; overconfidence; sensitivity analysis; significance testing

Abbreviations: CI, confidence interval; NHST, null-hypothesis significance testing; PBA, probabilistic bias analysis; RR, relative risk.

> "[T]here is no shame in not knowing. The problem arises when irrational thought and attendant behavior fill the vacuum left by ignorance."
>
> —Neil deGrasse Tyson (1, p. 38)

METHODOLOGY, LIKE SCIENCE, IS HYPOTHETICAL

In an accompanying article, Dr. Timothy Lash (2) describes how null-hypothesis significance testing (NHST) has contributed to problems of reproducibility, and discusses analytical methods for better capturing uncertainties of inference. These problems, however, are at least partly attributable to exclusive focus on random error and mechanistic biases in statistics while neglecting cognitive biases and other psychosocial factors affecting scientific inferences. Thus, in the present paper, I detail 3 cognitive distortions that are aggravated or induced by NHST: dichotomania, nullism, and reification.

To counter such cognitive problems of inference, the following methodological points need emphasis throughout teaching and research in health, medical, and social sciences:

1. The processes generating our observations are far too complex for us to capture all of their potentially important features, and their complete form is mostly beyond correct

intuitive understanding. Sophistication of a model does not mitigate misuse, however, for misuse becomes more opaque and tenaciously defensible when the model becomes harder to understand. Thus, analytical methodology at best provides frameworks for forcing some degree of logical consistency into inferential arguments, and examples of how these arguments can go wrong.

2. Methods do not come with real-world guarantees that they "work" in our application (get us closer to the truth than if we had ignored their outputs); theoretical "optimality" results are based on assumptions that are uncertain in reality. We are thus foolish if we take their uncertainty assessments (e.g., interval estimates) as sufficient for inference.
3. Inferences demand patient psychological as well as logical analysis, for our intuitions influence our judgments and in turn are heavily biased by our values, what we were taught, and what we have taught—however wrong those teachings are.
4. Statistical analyses are merely thought experiments, informing us as to what would follow deductively under their assumptions. These hypothetical experiments can train our intuitions but can also bias our inferences via *anchoring* (treating our primary analysis results as a specially upweighted reference point, even when there is no empirical basis for that) (3) and *reification* (acting as if our models are physical laws), as typifies rote statistical applications. These problems contribute to overconfident inference. Misinterpretations of statistical tests and their confinement to NHST are among the most prominent examples.
5. Any model that fits the data acceptably well will be only one of many possible data-generating mechanisms that we cannot rule out given our limited data and understanding. Sensitivity and bias analysis can help address this fundamental knowledge limitation but are in no way immune from cognitive distortions. If anything, they offer even more opportunities for misinterpretation and misuse, and may encourage overconfidence by appearing comprehensive.

I have discussed most of these points elsewhere (4–6), so I will focus on some specific problems raised by Lash's articles (2, 3) that seem neglected in most of the "replication crisis" literature, along with some limitations of sensitivity and bias analysis in addressing these problems.

I argue that current training in statistics and analytical methods is inadequate for addressing major sources of inference distortion, and that it should be expanded to cover the biased perceptual and thinking processes (cognitive biases) that plague research reports. As commonly misused, null-hypothesis significance testing (NHST) combines several cognitive problems to create highly distorted interpretations of study results. Interval estimation has proven highly vulnerable to the same problems. Sensitivity and bias analyses address model uncertainties by varying and relaxing assumptions, but (like Bayesian analyses) they are difficult to perform with proper accounting for prior information and are easily manipulated because they depend on specification of many models and parameters. Surprisingly, *P* values can be reframed to lessen cognitive problems by 1) presenting them without reference to a cutoff, 2) providing them for relevant alternatives to the null hypothesis, and 3) interpreting them with reference to all assumptions used in their computation rather than just the parameter they are tailored to test. *P* values, however, are poorly scaled for measuring evidence, a problem which could be addressed by transforming them into the information they supply against the model used to compute them.

THE NHST PROBLEM ARISES FROM A SYNERGY OF DICHOTOMANIA AND NULLISM

In his article, Lash (2) gives a telling account of literature distortions caused by NHST. After the publication of hundreds of papers and books explaining NHST problems over the past 75 years (e.g., see the citations in Greenland et al. (7)), it is indeed disheartening that NHST and its variants remain at the core of most analyses, apart from the relatively few journals that discourage statistical tests.

Those journals have usually requested the use of confidence intervals instead. Has forcing replacement of testing with confidence intervals addressed the problems that arose from NHST? As Lash explains (2), not as much as hoped. That should be unsurprising, because both confidence intervals and α-level tests were conceived as decision rules for behavior (8) but were rapidly misinterpreted as rules for belief, and thus fed the false notion that a single study can by itself tell us whether an effect is present or absent. They do so by degrading continuous measures of evidence into decisive conclusions, feeding the strong cognitive bias of *dichotomania*: the compulsion to replace quantities with dichotomies ("black-and-white thinking"), even when such dichotomization is unnecessary and misleading for inference.

As has long been known (9–11), use of the term "significant" or dichotomization of *P* values by comparing them with a fixed cutoff serves no good purpose for inference—it is less misleading and more informative to say (for example) that an association had a *P* value of 0.02 instead of "was significant" or had a *P* value of 0.17 instead of "was not significant" (12). Degrading *P* values and confidence intervals into null tests blinds the user to actual data patterns (13), thus invalidating conclusions and sometimes rendering them ludicrous. In a sadly typical example, one research group claimed that their study findings conflicted with earlier results because their estimated risk ratio was 1.20 (95% confidence interval: 0.97, 1.48) as opposed to a previously reported risk ratio of 1.20 (95% confidence interval: 1.09, 1.33) (14). Such idiocies are easy to find (15, Figure 3; 16; 17, p. 161; 18) and may be why one journal banned the use of confidence intervals along with statistical tests (19).

The distortion of focusing on the null value instead of the entire confidence interval dovetails too well with pressures to make results sound decisive. This null obsession is the most destructive pseudoscientific gift that conventional statistics (both frequentist and Bayesian) has given the modern world. One of its many damaging manifestations is *nullism* (also known as pseudo-skepticism): a religious faith that nature graces us with null associations in most settings. This faith should always be challenged within the applied context.

Instead, it goes unnoticed in the vast majority of education and practice—often to great harm.

Nullism appears to be a bias in science culture stemming from ostensibly "skeptical" scientific attitudes, along with rational desires to avoid false leads; it has been formalized in statistical tests designed to counter natural tendencies to see patterns in noise. The bias is built directly into Bayesian hypothesis testing in the form of spikes of prior probability placed on null hypotheses. Yet in soft sciences these spikes rarely have any basis in (and often conflict with) actual prior information (20–25). Medical research provides typical examples: Drugs and devices are approved precisely because of evidence that they affect human physiology, making the null hypothesis of no side effects *less* likely than some alternatives (22).

In frequentist hypothesis testing, nullism manifests itself as an implicit default assumption that false-positive inferences are always far more costly than false-negative ones. This in turn leads to adoption of test criteria that minimize false-positive rates no matter how many true effects are missed, and retardation of the process of scientific discovery (26). Neyman himself recognized that nullism is an incorrect general view, noting that false negatives could be more costly than false positives for some stakeholders (27, pp. 104–108; 28). Consider adverse drug effects: For the drug manufacturer, a false-negative inference can be far less costly than a false-positive one. Standard study-design criteria assume this cost difference with the requirement of a 5% maximum false-positive (type I error) rate and 80% minimum power, corresponding to a 20% maximum false-negative (type II error) rate and an implicit prior probability that adverse effects are unlikely. Yet, for a patient receiving the drug, the cost of a false-negative inference can be far higher (e.g., death or disability) than the cost of a false-positive one (e.g., having to use another drug). Thus, in hazard assessment, the traditional focus on testing only the null hypothesis is biased in favor of those who would be found liable for harms. This null bias is increased dramatically by multiple-comparison adjustments, which preserve false-positive rates at the expense of inflated false-negative rates, without regard to error costs or prior probabilities.

Some null-biased procedures (such as shrinkage methods) do have justifications in certain contexts, such as model selection and exploration; genomics provides examples with biological arguments for expecting few nonnegligible effects, along with a need to drastically reduce the number of associations pursued. Elsewhere, however, nullism seems to reflect a basic human aversion to admitting ignorance and uncertainty: Rather than recognize and explain why available evidence is inconclusive, experts freely declare that "the scientific method" treats the null as true until it is proven false, which is nothing more than a fallacy favoring those who benefit from belief in the null (29). Worse, this bias is often justified with wishful biological arguments (e.g., that we miraculously evolved toxicological defenses that can handle all modern chemical exposures) and basic epistemic mistakes—notably, thinking that parsimony is a property of nature when it is instead only an effective learning heuristic (30), or that refutationism involves believing hypotheses until they are falsified, when instead it involves never asserting a hypothesis is true (31).

Interval estimation could have addressed these problems had it been treated as its proponents advised: by careful examination and discussion of the full range of the interval and its vicinity to see what uncertainty would remain even if there were no validity problems, rather than focusing on whether it contained the null. Alas, this did not happen, and after generations of pleas for the use of confidence intervals (9, 10, 32, 33), we still see them being used to encourage dichotomous thinking (inside the interval vs. outside), nullism (by examining only whether the null value is within the interval), and overconfident inferences (as their name encourages).

It seems unappreciated that *P* values can help address these problems *if* they are computed for relevant nonnull hypotheses ("alternatives") as well as the null. For example, it is often claimed that a study provided evidence against an effect because the null test was "nonsignificant" with high power; that claim is revealed as wrong and deceptive when the test of an important alternative is even less significant (34). This information is supplied by a *P* value function (or confidence distribution) (15, 17, 33, 35), which provides *P* values for a full range of hypotheses and confidence intervals for a full range of confidence levels—thus addressing the criticism that null *P* values confound effect size with statistical precision (36). The *P* value function, or at least presentation of *P* values for effect sizes other than the null, can thus rescue the *P* value concept from the abuses inherent in NHST.

One-sided *P* values can further help mitigate nullism by shifting the focus from a precise hypothesis (such as the null), which is unlikely to be exactly true, to the hypothesis or probability that the targeted parameter lies in a particular direction (23, 37). Confidence intervals remain valuable, but only if they are interpreted to indicate the uncertainty or precision of the estimates under the model used to compute them (38, 39).

INTERPRETATIONS OF *P* VALUES AND CONFIDENCE INTERVALS IN A WORLD OF BIAS

Even if we draw a *P* value function, there remains the problem of properly interpreting the *P* values it provides (7). This problem is compounded when assumptions used in the analysis have not been enforced by the design and conduct of the study (40). For example, assumptions of "no unmeasured confounding" and "conditionally ignorable treatment assignment" are operationally equivalent to claiming that our data were produced by some kind of intricately designed randomized experiment, and thus (by definition) are not enforced and are often doubtful in observational research (23, 41). And the usual distributional assumptions of statistics can be severely violated whenever analysis decisions are not captured in the analysis model (40, 42).

Sensitivity to plausible assumption violations (model dependence) is a major underappreciated weakness of all reasoning. Even so-called "robust" statistical methods are sensitive to assumption violations represented by uncontrolled biases. These violations should be expected in human-subjects research and render hypothetical any formal statistical inferences about causation (6). Confronting this reality, one way to make sense of conventional statistics is to reorient our interpretations to be *unconditional* on model assumptions: Instead of thinking of a *P* value or confidence interval as referring to a single parameter (such as a model coefficient), we can think of it as referring to the

entire model it was computed from, including all assumptions about bias (especially implicit assumptions) (7).

Traditionally, coefficient tests are taken to refer only to the assumption that the coefficient equals the tested value, given all the other model assumptions. This tradition is pernicious whenever (as is always the case in soft sciences) the other model assumptions are far from guaranteed: All inferential statistics (whether *P* values, confidence intervals, likelihood ratios, or posterior probabilities) are heavily influenced by violations of validity assumptions arising from uncontrolled nonlinearity, confounding, measurement error, selection bias, *P*-hacking, or fraud. Because almost all assumptions are uncertain, a small *P* value only signals that there may be a problem with at least 1 assumption, without saying which one. Asymmetrically, a large *P* value only means that this particular test did not detect a problem—perhaps because there is none, or because the test is insensitive to the problems, or because biases and random errors largely canceled each other out. We recognize these possibilities when we admit that results (whether with small or large *P* values) may be "due to chance *or* bias."

Uncertainty about validity assumptions is not captured by standard testing descriptions—in fact, assumption uncertainty is a core weakness of conventional statistics, which depends on reification to connect its outputs to the real world. This weakness can be addressed by recognizing that a *P* value does *not* test only 1 hypothesis if the other assumptions are uncertain. Rather, it is a test of *every* assumption used to compute the test (24, p. 75). For example, a so-called null test is really a test of a model comprising all assumptions used to compute the *P* value, including validity assumptions as well as the null hypothesis. This is so even if the test is tailored hypothetically to have "high power" for the targeted parameter (i.e., derived to maximize power to detect violations along the particular dimension specified by the null hypothesis).

P VALUES AND EVIDENCE MEASURES

Although Bayesians have raised important criticisms of significance testing, they often overlook limitations of Bayesian inference (43, 44) and sometimes claim that *P* values overstate evidence against the null (45–47). That claim is mistaken insofar as it blames the *P* value for misinterpretations by teachers and users of statistics; furthermore, it is based on a Bayesian standard of evidence (the Bayes factor) which is of doubtful validity for evaluating refutational measures like the frequentist *P* value (20, 48).

A genuine cognitive problem is that a *P* value forces the test statistic into the unit (0–1) scale, which renders it a highly nonlinear and nonintuitive function of data information. One way to address this problem is to treat a *P* value not as an evidence measure but instead as merely an index of compatibility between the test statistic and the model (set of all assumptions) used to compute the *P* value, on a scale of 0 to 1, where 0 = completely incompatible (statistic impossible under the model) and 1 = completely compatible (statistic exactly as predicted by the model) (7). The refutational strength of a *P* value, however, can be gauged by translating it into the bits of information it supplies against the model. For a *P* value of p, this quantity is $-\log_2(p)$, called the *surprisal* (49) in seeing an event of probability p if the model is correct. This measure is 0 (unsurprising) when $P = 1$, and it increases exponentially as *P* declines. The number of bits of information against the model supplied by $P = 0.05$ is then only $-\log_2(0.05) = 4.3$; this is about as surprising as seeing 4 heads in 4 fair coin tosses, which has a probability of $1/2^4 = 0.0625$, thus conveying $-\log_2(1/2^4) = 4$ bits of information against fairness of the tosses. For comparison, $P = 0.01$ and $P = 0.09$ translate to $-\log_2(0.01) = 6.6$ and $-\log_2(0.09) = 3.5$. Thus, any evidence overstatement lies not with the *P* value but with 0.05-dichotomaniacs who mistakenly think that $P = 0.05$ represents just enough evidence to reject the model, instead of recognizing it as a small amount of evidence against the model.

BEYOND CONVENTIONAL STATISTICS: THE PERILOUS QUEST FOR REALISTIC AND RELEVANT METHODS

To place sensitivity and bias analyses in the generalized-model framework described above, consider an adjusted relative risk (RR) parameter RR_{adj} as estimated by the usual sort of risk regression, propensity scoring, or some combination (such as doubly robust regression). Conventional statistics only refer to RR_{adj} because that is all one can identify without introducing external ("prior") information about the function connecting it to the targeted causal relative risk RR_{causal}. In methodology, this profound knowledge gap is usually dealt with by saying that the statistics refer to RR_{causal} conditional on the adjustments being sufficient to remove bias. This treatment dodges the fact that RR_{adj} is actually a complex, unknown function of the target effect RR_{causal}, the data, and various unknown bias parameters, so that tests and estimates of RR_{adj} omit major sources of uncertainty about the effect RR_{causal} and by themselves place no limit on its size.

Ideally, study-design features would identify the bias function or even force RR_{adj} to equal RR_{causal}, but nothing so ambitious can be achieved in typical observational studies. Inferences derived from statistical analysis may nonetheless appear compelling simply because they are plausible in light of what is known. This plausibility may lull one into forgetting that other analyses may fit the same data equally well using plausible but very different assumptions about the bias function, and thus lead to very different inferences. In the philosophy of science, this logical limit of knowledge is known as the underdetermination of scientific theories by observations (50), and it corresponds to statistical nonidentification of the bias function linking RR_{adj} to RR_{causal}.

Statistics traditionally deals with this problem by forcing identification of RR_{causal} using some conventional model without worrying too much about whether the model is remotely plausible, instead appealing to insensitive tests of fit. Bias analysis tries to reintroduce plausibility by estimating the function connecting RR_{adj} to RR_{causal} from a combination of background information (such as validation studies), arbitrary specifications (such as distributional shapes and independencies), and what little data information there may be on residual bias. The assumptions introduced are hopefully less absurd than claiming $RR_{adj} = RR_{causal}$, but there is no guarantee that this is so (e.g., as with absurd assumptions that bias parameters are uniformly distributed or are independent between cases and controls).

Even with realistic choices, the sensitivity of sensitivity and bias analyses must be evaluated (51). The plausibility of an estimated bias function is determined by intuitions, prejudices, and understanding of the applied context; those can vary dramatically across researchers, in turn leading to very different specifications and inferences even if they are anchored to the same conventional analysis. Adding to this problem, sensitivity and bias analyses are more difficult to perform correctly and more easily massaged toward preferred conclusions, because they require specification of many more equations and their parameters. And unlike NHST, abuse of sensitivity and bias analysis is as yet barely studied because the pool of such analyses remains small and highly selective. It thus seems implausible that these analyses will increase replicability of inferences, although they can reveal how assumptions affect those inferences. (Here "replicability" is used according to recommendations of the American Statistical Association (52) to denote independent checks of reported results with new data; "reproducibility" then denotes checks of reported results using the original data and computer code.)

As with Bayesian statistical methods, probabilistic bias analysis (PBA)—including Bayesian bias analysis as well as probabilistic sensitivity analysis—is especially hazardous because of poor defaults and intuitions regarding prior distributions for parameters (53, pp. 369–372). One may thus doubt whether individual studies should go so far as a full PBA (53, pp. 347 and 380). Among the objections (which also apply to other sophisticated analysis methods):

1. No inference should be based on a single study alone, even if that study was designed to be the final input into a policy decision. Research synthesis is needed to reach reliable inferences, and that requires detailed methods and data descriptions for each study. It would thus be damaging if publications omitted such details in favor of PBA, which itself requires lengthy description.
2. Like any analysis, PBA is simply a thought experiment predicated on assumptions that may be in error, with outputs highly sensitive to those assumptions. But the sophistication of PBA may seduce users into making overconfident claims about the analysis results, and may increase anchoring of subsequent judgments to those results.
3. Researchers and referees have demonstrated severe problems in using basic ideas like *P* values and confidence intervals correctly. Should we expect fewer problems with sensitivity and bias analyses? Especially, PBA is an order of magnitude more subtle and complex, requiring integration of multiple uncertainty sources and models. Complex models increase the potential for oversights and hidden errors.
4. The unlimited sensitivity of effect estimates from bias models implies that any desired inference can be manufactured by back-calculating to the plausible-looking models and priors that produce it, thus providing an avenue for motivated statistical reasoning (54). Analysts can completely deceive readers (and themselves) by failing to report result-driven analysis selection.

A narrower concern is the relatively untested nature of PBA software. As an example, a bug in one meta-analytical PBA (55) was only discovered years later when a colleague attempted to reproduce the results using other software (Dr. Timothy Mak, University of Hong Kong, personal communication, 2010); fortunately, the correction did not alter the main inference that the studies being combined failed to establish anything (thus illustrating a major robustness advantage of ambiguous conclusions).

None of the above argues against the potential value of well-done, transparent PBA for research synthesis to inform decisions and policy. In fact, one can demand PBA in support of contestable claims about policy implications (53, pp. 347 and 380). But warnings against policy claims within single studies (56) extend to PBA: Like policy analysis, PBA remains a highly technical topic in its own right, demanding well-developed methods such as posterior sampling alongside as-yet-underdeveloped methods such as prior modeling (by which I do *not* mean prior elicitation, but rather extraction and coding of relevant information from other studies). Thus, as with policy analysis, the effort and detailed reporting needed for good PBA requires its own article, which may be hard to justify when conventional methods yield ambiguous results.

CONCLUSIONS

Viewing the distortions generated by conventional statistical teaching and practice, I see a dire need to get away from inferential statistics and hew more closely to descriptions of study procedures, data collection (which may have occurred before the study), and the resulting data. This recommendation runs against ambitions and pressures on authors to expound on the implications of their own studies, however biased and naive their exposition. But what science and society need most from a study is its data (or numerical summaries that allow adequate reconstruction of the data) and thorough documentation of how those data were generated, so that sources of uncertainty can be recognized and the study information can be accurately entered into research syntheses (57).

Instead, conventional statistical training seems to encourage human tendencies toward overconfidence and conclusiveness by providing numerically precise answers to hypothetical experiments and decision problems. The artificial problems that conventional statistics solves are often far removed from the actual research contexts in soft sciences like health and medicine. NHST is value-biased as well, with implicit loss functions that would be unacceptable to many stakeholders—if they were revealed (5, 8, 26–28). Decades of piecemeal objections to the resulting abuses have reduced distortions in epidemiology, but the core problems remain common in the broader literature.

I am thus unable to escape the inference that training in statistics and analytical methods has shown itself deficient in addressing major sources of inference distortion. We can begin to address this deficiency by adding overviews of the now-vast literature on cognitive biases and debiasing techniques (58–61) to basic statistics and methods courses (for 2 decades, I used a text by Gilovich (62), a $10 paperback, in my course on logic, causation, and probability; a Web search on "cognitive biases" will reveal many up-to-date nontechnical treatments of the topic (63–66)). We also need to investigate how cognitive biases have affected research literature. Methodologists should formulate these teaching and research programs

collaboratively with experts in cognitive sciences, social psychology, and behavioral economics, paying special attention to biases in methodology as well as in reported inferences.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California (Sander Greenland); and Department of Statistics, College of Letters and Science, University of California, Los Angeles, Los Angeles, CA (Sander Greenland).

I am grateful to Drs. Steve Cole, David Krantz, Timothy Lash, Robert Lyles, and Raymond Neutra for helpful comments on this paper. Any errors that remain are solely my responsibility.

Conflict of interest: none declared.

REFERENCES

1. Tyson ND. *The Sky Is Not the Limit: Adventures of an Urban Astrophysicist*. 2nd ed. Amherst, NY: Prometheus Books; 2004.
2. Lash TL. The harm done to reproducibility by the culture of null hypothesis significance testing. *Am J Epidemiol*. 2017; 186(6):627–635.
3. Lash TL. Heuristic thinking and inference from observational epidemiology. *Epidemiology*. 2007;18(1):67–72.
4. Greenland S. Causal inference as a prediction problem: assumptions, identification, and evidence synthesis. In: Berzuini C, Dawid AP, Bernardinelli L, eds. *Causality Inference: Statistical Perspectives and Applications*. Chichester, United Kingdom: John Wiley & Sons Ltd.; 2012:43–58.
5. Greenland S. Transparency and disclosure, neutrality and balance: shared values or just shared words? *J Epidemiol Community Health*. 2012;66(11):967–970.
6. Greenland S. For and against methodology: some perspectives on recent causal and statistical inference debates. *Eur J Epidemiol*. 2017;32(1):3–20.
7. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337–350.
8. Neyman J. "Inductive behavior" as a basic concept of philosophy of science. *Rev Int Stat Inst*. 1957;25(1/3):7–22.
9. Rothman KJ. A show of confidence. *N Engl J Med*. 1978; 299(24):1362–1363.
10. Rothman KJ. Significance questing. *Ann Intern Med*. 1986; 105(3):445–447.
11. Rothman KJ. Curbing type I and type II errors. *Eur J Epidemiol*. 2010;25(4):223–224.
12. Amrhein V, Korner-Nievergelt F, Roth T. The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. *PeerJ*. 2017;5:e3544.
13. McShane BB, Gal D. Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Manag Sci*. 2016;62(6):1707–1718.
14. Schmidt M, Rothman KJ. Mistaken inference caused by reliance on and misinterpretation of a significance test. *Int J Cardiol*. 2014;177(3):1089–1090.
15. Poole C. Beyond the confidence interval. *Am J Public Health*. 1987;77(2):195–199.
16. Rothman KJ, Lanes S, Robins J. Causal inference (letter). *Epidemiology*. 1993;4(6):555–556.
17. Rothman KJ, Greenland S, Lash TL. Precision and statistics in epidemiologic studies. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008:148–167.
18. Greenland S. A serious misinterpretation of a consistent inverse association of statin use with glioma across 3 case-control studies. *Eur J Epidemiol*. 2017;32(1):87–88.
19. Trafimow D, Marks M. Editorial. *Basic Appl Soc Psych*. 2015; 37(1):1–2.
20. Casella G, Berger RL. Comment. *Stat Sci*. 1987;2(3):344–347.
21. Greenland S. Weaknesses of certain Bayesian methods for meta-analysis: the case of vitamin E and mortality. *Clin Trials*. 2009;6:42–46.
22. Greenland S. Null misinterpretation in statistical testing and its impact on health risk assessment. *Prev Med*. 2011;53(4–5): 225–228.
23. Greenland S, Poole C. Living with *p* values: resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology*. 2013;24(1):62–68.
24. Greenland S, Poole C. Living with statistics in observational research. *Epidemiology*. 2013;24(1):73–78.
25. Gelman A. *P* values and statistical practice. *Epidemiology*. 2013;24(1):69–72.
26. Fiedler K, Kutzner F, Krueger JI. The long way from α-error control to validity proper: problems with a short-sighted false-positive debate. *Perspect Psychol Sci*. 2012;7(6): 661–669.
27. Neyman J. Frequentist probability and frequentist statistics. *Synthese*. 1977;36(1):97–131.
28. Greenland S. The ASA guidelines and null bias in current teaching and practice. *Am Statist*. 2016;70(suppl 10). http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/utas_a_1154108_sm5079.pdf. Accessed July 21, 2017.
29. Greenland S. The need for critical appraisal of expert witnesses in epidemiology and statistics. *Wake Forest Law Rev*. 2004;39: 291–310.
30. Kelly KT. Simplicity, truth, and probability. In: Bandyopadhyay PS, Forster MR, eds. *Philosophy of Statistics*. 1st ed. (Handbook of the Philosophy of Science, vol. 7). Amsterdam, the Netherlands: Elsevier B.V.; 2011:983–1026.
31. Popper KR. *The Logic of Scientific Discovery*. New York, NY: Basic Books; 1959.
32. Yates F. The influence of statistical methods for research workers on the development of the science of statistics. *J Am Stat Assoc*. 1951;46(253):19–34.
33. Cox DR. Some problems connected with statistical inference. *Ann Math Stat*. 1958;29(2):357–372.
34. Greenland S. Nonsignificance plus high power does not imply support for the null over the alternative. *Ann Epidemiol*. 2012; 22(5):364–368.
35. Birnbaum A. A unified theory of estimation, I. *Ann Math Stat*. 1961;32:112–135.
36. Lang JM, Rothman KJ, Cann CI. That confounded *P*-value. *Epidemiology*. 1998;9(1):7–8.
37. Casella G, Berger RL. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J Am Stat Assoc*. 1987;82(397):106–111.
38. Poole C. Confidence intervals exclude nothing. *Am J Public Health*. 1987;77(4):492–493.

39. Poole C. Low *P*-values or narrow confidence intervals: which are more durable? *Epidemiology*. 2001;12(3):291–294.
40. Gelman A. The problems with *p*-values are not just with *p*-values: my comments on the recent ASA statement (with comments) [blog post]. http://andrewgelman.com/2016/03/07/29212/. Accessed May 21, 2017.
41. Greenland S. Randomization, statistics, and causal inference. *Epidemiology*. 1990;1(6):421–429.
42. Gelman A, Loken E. The statistical crisis in science: data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *Am Sci*. 2014;102(6):460–465.
43. Kelly KT, Glymour C. Why probability does not capture the logic of scientific justification. In Hitchcock C, ed. *Contemporary Debates in the Philosophy of Science*. London, United Kingdom: Blackwell Ltd.; 2004.
44. Gigerenzer G, Marewski JN. Surrogate science: the idol of a universal method for scientific inference. *J Manag*. 2015;41(2):421–440.
45. Sellke TM, Bayarri MJ, Berger JO. Calibration of *p* values for testing precise null hypotheses. *Am Stat*. 2001;55(1):62–71.
46. Goodman SN. Introduction to Bayesian methods I: measuring the strength of evidence. *Clin Trials*. 2005;2(4):282–290.
47. Wagenmakers EJ. A practical solution to the pervasive problem of *p* values. *Psychon Bull Rev*. 2007;14(5):779–804.
48. Senn S. A comment on replication, *p*-values and evidence. S.N. Goodman, *Statistics in Medicine* 1992;11:875–879 [letter]. *Stat Med*. 2002;21(16):2437–2444.
49. Fraundorf P. Examples of surprisal. 2017. http://www.umsl.edu/~fraundorfp/egsurpri.html. Accessed April 6, 2017.
50. Stanford K. Underdetermination of scientific theory. In: Zalta EN, ed. *Stanford Encyclopedia of Philosophy Archive. Spring 2016 Edition*. https://plato.stanford.edu/archives/spr2016/entries/scientific-underdetermination/. Published August 12, 2009. Revised September 16, 2013. Accessed April 6, 2017.
51. Greenland S. The sensitivity of a sensitivity analysis (invited paper). In: *1997 Proceedings of the Biometrics Section*. Alexandria, VA: American Statistical Association; 1998:19–21.
52. Broman K, Cetinkaya-Rundel M, Nussbaum A, et al. *Recommendations to Funding Agencies for Supporting Reproducible Research*. Washington, DC: American Statistical Association; 2017. http://www.amstat.org/asa/files/pdfs/pol-reproducibleresearchrecommendations.pdf. Accessed April 6, 2017.
53. Greenland S, Lash TL. Bias analysis. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008:345–380.
54. Kahan DM, Peters E, Dawson EC, et al. Motivated numeracy and enlightened self-government. *Behav Public Pol*. 2017;1(1):54–86.
55. Greenland S, Kheifets L. Leukemia attributable to residential magnetic fields: results from analyses allowing for study biases. *Risk Anal*. 2006;26(2):471–482.
56. Rothman KJ. Policy recommendations in epidemiologic research papers. *Epidemiology*. 1993;4(2):94–95.
57. Greenland S, Gago-Domiguez M, Castelao JE. The value of risk-factor ("black-box") epidemiology. *Epidemiology*. 2004;15(5):529–535.
58. Kahneman D, Slovic P, Tversky P, eds. *Judgment Under Uncertainty: Heuristics and Biases*. New York, NY: Cambridge University Press; 1982.
59. Gilovich T, Griffin D, Kahneman D. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York, NY: Cambridge University Press; 2002.
60. Baron J. *Thinking and Deciding*. 4th ed. New York, NY: Cambridge University Press; 2007.
61. Pohl RF, ed. *Cognitive Illusions: Intriguing Phenomena in Thinking, Judgment and Memory*. 2nd ed. New York, NY: Routledge, Taylor & Francis Group LLC; 2017.
62. Gilovich T. *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life*. New York, NY: Free Press; 1993.
63. Kahneman D. *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux; 2011.
64. LessWrongWiki. Bias [blog post]. https://wiki.lesswrong.com/wiki/Bias. Modified November 19, 2013. Accessed April 25, 2017.
65. Wikipedia. Cognitive bias. https://en.wikipedia.org/wiki/Cognitive_bias. Modified July 3, 2017. Accessed April 24, 2017.
66. Wikipedia. List of cognitive biases. https://en.wikipedia.org/wiki/List_of_cognitive_biases. Modified July 17, 2017. Accessed April 24, 2017.