Statistical Tests, *P*-values, Confidence Intervals, and Power: A Guide to Misinterpretations

Sander GREENLAND, Stephen J. SENN, Kenneth J. ROTHMAN, John B. CARLIN, Charles POOLE, Steven N. GOODMAN, and Douglas G. ALTMAN

Misinterpretation and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof. Instead, correct use and interpretation of these statistics requires an attention to detail which seems to tax the patience of working scientists. This high cognitive demand has led to an epidemic of shortcut definitions and interpretations that are simply wrong, sometimes disastrously so—and yet these misinterpretations dominate much of the scientific literature.

In light of this problem, we provide definitions and a discussion of basic statistics that are more general and critical than typically found in traditional introductory expositions. Our goal is to provide a resource for instructors, researchers, and consumers of statistics whose knowledge of statistical theory and technique may be limited but who wish to avoid and spot misinterpretations. We emphasize how violation of often unstated analysis protocols (such as selecting analyses for presentation based on the *P*-values they produce) can lead to small *P*-values even if the declared test hypothesis is correct, and can lead to large *P*-values even if that hypothesis is incorrect. We then provide an explanatory list of 25 misinterpretations of *P*-values, confidence intervals, and power. We conclude with guidelines for improving statistical interpretation and reporting.

KEY WORDS: Confidence intervals; Hypothesis testing; Null testing; *P*-value; Power; Significance tests; Statistical testing.

Introduction

Misinterpretation and abuse of statistical tests has been decried for decades, yet remains so rampant that some scientific journals discourage use of "statistical significance" (classifying results as "significant" or not based on a *P*-value) (Lang et al. 1998). One journal now bans all statistical tests and mathematically related procedures such as confidence intervals (Trafimow and Marks 2015), which has led to considerable discussion and debate about the merits of such bans (e.g., Ashworth 2015; Flanagan 2015).

Despite such bans, we expect that the statistical methods at issue will be with us for many years to come. We thus think it imperative that basic teaching as well as general understanding of these methods be improved. Toward that end, we attempt to explain the meaning of significance tests, confidence intervals, and statistical power in a more general and critical way than is traditionally done, and then review 25 common misconceptions in light of our explanations. We also discuss a few more subtle but nonetheless pervasive problems, explaining why it is important to examine and synthesize all results relating to a scientific question, rather than focus on individual findings. We further explain why statistical tests should never constitute the sole input to inferences or decisions about associations or effects. Among the many reasons are that, in most scientific settings, the arbitrary classification of results into "significant" and "nonsignificant" is unnecessary for and often damaging to valid interpretation of data; and that estimation of the size of effects and the uncertainty surrounding our estimates will be far more important for scientific inference and sound judgment than any such classification.

More detailed discussion of the general issues can be found in many articles, chapters, and books on statistical methods and their interpretation (e.g., Altman et al. 2000; Atkins and Jarrett 1979; Cox 1977, 1982; Cox and Hinkley 1974; Freedman et al. 2007; Gibbons and Pratt 1975; Gigerenzer et al. 1990, Ch. 3; Harlow et al. 1997; Hogben 1957; Kaye and Freedman 2011; Morrison and Henkel 1970; Oakes 1986; Pratt 1965; Rothman et al. 2008, Ch. 10; Ware et al. 2009; Ziliak and McCloskey 2008). Specific issues are covered at length in these sources and in the many peer-reviewed articles that critique common misinterpretations of null-hypothesis testing and "statistical significance" (e.g., Altman and Bland 1995; Anscombe 1990; Bakan 1966; Bandt and Boen 1972; Berkson 1942; Bland and Altman 2015; Chia 1997; Cohen 1994; Evans et al. 1988; Fidler and Loftus 2009; Gardner and Altman 1986; Gelman 2013; Gelman and Loken 2014; Gelman and Stern 2006; Gigerenzer 2004;

Gigerenzer and Marewski 2015; Goodman 1992, 1993, 1999, 2008; Greenland 2011, 2012ab; Greenland and Poole, 2011, 2013ab; Grieve 2015; Harlow et al. 1997; Hoekstra et al. 2006; Hurlbert and Lombardi 2009; Kaye 1986; Lambdin 2012; Lang et al. 1998; Langman 1986; LeCoutre et al. 2003; Lew 2012; Loftus 1996; Matthews and Altman 1996a; Pocock and Ware 2009; Pocock et al. 1987; Poole 1987ab, 2001; Rosnow and Rosenthal 1989; Rothman 1978, 1986; Rozeboom 1960; Salsburg 1985; Schmidt 1996; Schmidt and Hunter 2002; Sterne and Davey Smith 2001; Thompson 1987; Thompson 2004; Wagenmakers 2007; Walker 1986; Wood et al. 2014).

Statistical Tests, *P*-values, and Confidence Intervals: A Caustic Primer

Statistical Models, Hypotheses, and Tests

Every method of statistical inference depends on a complex web of assumptions about how data were collected and analyzed, and how the analysis results were selected for presentation. The full set of assumptions is embodied in a *statistical model* that underpins the method. This model is a mathematical representation of data variability, and thus ideally would capture accurately all sources of such variability. Many problems arise, however, because this statistical model often incorporates unrealistic or at best unjustified assumptions. This is true even for so-called "nonparametric" methods, which (like other methods) depend on assumptions of random sampling or randomization. These assumptions are often deceptively simple to write mathematically, yet in practice are difficult to satisfy and verify, as they may depend on successful completion of a long sequence of actions (such as identifying, contacting, obtaining consent from, obtaining cooperation of, and following up subjects, as well as adherence to study protocols for treatment allocation, masking, and data analysis).

There is also a serious problem of defining the scope of a model, in that it should allow not only for a good representation of the observed data but also of hypothetical alternative data that might have been observed. The reference frame for data that "might have been observed" is often unclear, for example if multiple outcome measures or multiple predictive factors have been measured, and many decisions surrounding analysis choices have been made after the data were collected—as is invariably the case (Gelman and Loken 2014).

The difficulty of understanding and assessing underlying assumptions is exacerbated by the fact that the statistical model is usually presented in a highly compressed and abstract form—if presented at all. As a result, many assumptions go unremarked and are often unrecognized by users as well as consumers of statistics. Nonetheless, all statistical methods and interpretations are premised on the model assumptions; that is, on an assumption that the model provides a valid representation of the variation we would expect to see across data sets, faithfully reflecting the circumstances surrounding the study and phenomena occurring within it.

In most applications of statistical testing, one assumption in the model is a hypothesis that a particular effect has a specific size, and has been targeted for statistical analysis. (For simplicity, we use the word "effect" when "association or effect" would arguably be better in allowing for noncausal studies such as most surveys.) This targeted assumption is called the study hypothesis or *test hypothesis*, and the statistical methods used to evaluate it are called *statistical hypothesis tests*. Most often, the targeted effect size is a "null" value representing *zero* effect (e.g., that the study treatment makes no difference in average outcome), in which case the test hypothesis is called the *null hypothesis*. Nonetheless, it is also possible to test other effect sizes. We may also test hypotheses that the effect does or does not fall within a specific range; for example, we may test the hypothesis that the effect is no greater than a particular amount, in which case the hypothesis is said to be a *one-sided* or *dividing* hypothesis (Cox 1977, 1982).

Much statistical teaching and practice has developed a strong (and unhealthy) focus on the idea that the main aim of a study should be to test null hypotheses. In fact most descriptions of statistical testing focus *only* on testing null hypotheses, and the entire topic has been called "Null Hypothesis Significance Testing" (NHST). This exclusive focus on null hypotheses contributes to misunderstanding of tests. Adding to the misunderstanding is that many authors (including R.A. Fisher) use "null hypothesis" to refer to any test hypothesis, even though this usage is at odds with other authors and with ordinary English definitions of "null"—as are statistical usages of "significance" and "confidence."

Uncertainty, Probability, and Statistical Significance

A more refined goal of statistical analysis is to provide an evaluation of certainty or uncertainty regarding the size of an effect. It is natural to express such certainty in terms of "probabilities" of hypotheses. In conventional statistical methods, however, "probability" refers not to hypotheses, but to quantities that are hypothetical frequencies of data patterns under an assumed statistical model. These methods are thus called *frequentist* methods, and the hypothetical frequencies they predict are called "frequency probabilities." Despite considerable training to the contrary, many statistically educated scientists revert to the habit of misinterpreting these frequency probabilities as hypothesis probabilities. (Even more confusingly, the term "likelihood of a parameter value" is reserved by statisticians to refer to the probability of the observed data *given* the parameter value; it does not refer to a probability of the parameter taking on the given value.)

Nowhere are these problems more rampant than in applications of a hypothetical frequency called the *P-value*, also known as the "observed significance level" for the test hypothesis. Statistical "significance tests" based on this concept have been a central part of statistical analyses for centuries (Stigler 1986). The focus of traditional definitions of *P*-values and statistical significance has been on null hypotheses, treating all other assumptions used to compute the *P*-value as if they were known to be correct. Recognizing that these other assumptions are often questionable if not unwarranted, we will adopt a more general view of the *P*-value as a statistical summary of the compatibility between the observed data and what we would predict or expect to see if we knew the entire statistical model (*all* the

assumptions used to compute the P-value) were correct.

Specifically, the distance between the data and the model prediction is measured using a *test statistic* (such as a t-statistic or a chi-squared statistic). The P-value is then the probability that the chosen test statistic would have been *at least* as large as its observed value if *every* model assumption were correct, including the test hypothesis. This definition embodies a crucial point lost in traditional definitions: In logical terms, the P-value tests *all* the assumptions about how the data were generated (the entire model), not just the targeted hypothesis it is supposed to test (such as a null hypothesis). Furthermore, these assumptions include far more than what are traditionally presented as modeling or probability assumptions—they include assumptions about the conduct of the analysis, for example that intermediate analysis results were not used to determine which analyses would be presented.

It is true that the smaller the P-value, the more unusual the data would be *if* every single assumption were correct; but a very small P-value does *not* tell us which assumption is incorrect. For example, the P-value may be very small because the targeted hypothesis is false; but it may instead (or in addition) be very small because the study protocols were violated, or because it was selected for presentation based on its small size. Conversely, a large P-value indicates only that the data are not unusual under the model, but does not imply that the model or any aspect of it (such as the targeted hypothesis) is correct; it may instead (or in addition) be large because (again) the study protocols were violated, or because it was selected for presentation based on its large size.

The general definition of a P-value may help one to understand why statistical tests tell us much less than what many think they do: Not only does a P-value *not* tell us whether the hypothesis targeted for testing is true or not; it says nothing specifically related to that hypothesis unless we can be completely assured that every other assumption used for its computation is correct—an assurance that is lacking in far too many studies.

Nonetheless, the P-value can be viewed as a continuous measure of the compatibility between the data and the entire model used to compute it, ranging from 0 for complete incompatibility to 1 for perfect compatibility, and in this sense may be viewed as measuring the fit of the model to the data. Too often, however, the P-value is degraded into a dichotomy in which results are declared "statistically significant" if P falls on or below a cut-off (usually 0.05) and declared "nonsignificant" otherwise. The terms "significance level" and "alpha level" (α) are often used to refer to the cut-off; however, the term "significance level" invites confusion of the cut-off with the P-value itself. Their difference is profound: the cut-off value α is supposed to be fixed in advance and is thus part of the study design, unchanged in light of the data. In contrast, the P-value is a number computed from the data and thus an analysis result, unknown until it is computed.

Moving From Tests to Estimates

We can vary the test hypothesis while leaving other assumptions unchanged, to see how the P-value differs across competing test hypotheses. Usually, these test hypotheses specify dif-ferent sizes for a targeted effect; for example, we may test the hypothesis that the average difference between two treatment groups is zero (the null hypothesis), or that it is 20 or –10 or any size of interest. The effect size whose test produced $P = 1$ is the size most compatible with the data (in the sense of predicting what was in fact observed) *if* all the other assumptions used in the test (the statistical model) were correct, and provides a *point estimate* of the effect under those assumptions. The effect sizes whose test produced $P > 0.05$ will typically define a range of sizes (e.g., from 11.0 to 19.5) that would be considered more compatible with the data (in the sense of the observations being closer to what the model predicted) than sizes outside the range—again, if the statistical model were correct. This range corresponds to a $1 - 0.05 = 0.95$ or 95% *confidence interval*, and provides a convenient way of summarizing the results of hypothesis tests for many effect sizes. Confidence intervals are examples of *interval estimates*.

Neyman (1937) proposed the construction of confidence intervals in this way because they have the following property: If one calculates, say, 95% confidence intervals repeatedly *in valid applications*, 95% of them, on average, will contain (i.e., include or cover) the true effect size. Hence, the specified confidence level is called the *coverage probability*. As Neyman stressed repeatedly, this coverage probability is a property of a long sequence of confidence intervals computed from valid models, rather than a property of any single confidence interval.

Many journals now require confidence intervals, but most textbooks and studies discuss P-values only for the null hypothesis of no effect. This exclusive focus on null hypotheses in testing not only contributes to misunderstanding of tests and underappreciation of estimation, but also obscures the close relationship between P-values and confidence intervals, as well as the weaknesses they share.

What P-values, Confidence Intervals, and Power Calculations Don't Tell Us

Much distortion arises from basic misunderstanding of what P-values and their relatives (such as confidence intervals) do *not* tell us. Therefore, based on the articles in our reference list, we review prevalent P-value misinterpretations as a way of moving toward defensible interpretations and presentations. We adopt the format of Goodman (2008) in providing a list of misinterpretations that can be used to critically evaluate conclusions offered by research reports and reviews. Every one of the italicized statements in our list has contributed to statistical distortion of the scientific literature, and we add the emphatic "No!" to underscore statements that are not only fallacious but also not "true enough for practical purposes."

Common Misinterpretations of Single P-values

1. The P-value is the probability that the test hypothesis is true; for example, if a test of the null hypothesis gave $P = 0.01$, the null hypothesis has only a 1% chance of being true; if instead it gave $P = 0.40$, the null hypothesis has a 40% chance of being true.—No! The P-value *assumes* the test hypothesis is true—it is *not* a hypothesis probability and may be far from any

reasonable probability for the test hypothesis. The *P*-value simply indicates the degree to which the data conform to the pattern predicted by the test hypothesis and all the other assumptions used in the test (the underlying statistical model). Thus $P = 0.01$ would indicate that the data are not very close to what the statistical model (including the test hypothesis) predicted they should be, while $P = 0.40$ would indicate that the data are much closer to the model prediction, allowing for chance variation.

2. *The P-value for the null hypothesis is the probability that chance alone produced the observed association; for example, if the P-value for the null hypothesis is 0.08, there is an 8% probability that chance alone produced the association.*—No! This is a common variation of the first fallacy and it is just as false. To say that chance *alone* produced the observed association is logically equivalent to asserting that every assumption used to compute the *P*-value is correct, *including the null hypothesis*. Thus to claim that the null *P*-value is the probability that chance alone produced the observed association is completely backwards: The *P*-value is a probability computed *assuming* chance was operating alone. The absurdity of the common backwards interpretation might be appreciated by pondering how the *P*-value, which is a probability deduced *from* a set of assumptions (the statistical model), can possibly refer to the probability *of* those assumptions.

Note: One often sees "alone" dropped from this description (becoming "the *P*-value for the null hypothesis is the probability that chance produced the observed association"), so that the statement is more ambiguous, but just as wrong.

3. *A significant test result (P ≤ 0.05) means that the test hypothesis is false or should be rejected.*—No! A small *P*-value simply flags the data as being unusual if all the assumptions used to compute it (including the test hypothesis) were correct; it may be small because there was a large random error or because some assumption other than the test hypothesis was violated (for example, the assumption that this *P*-value was not selected for presentation because it was below 0.05). $P \leq 0.05$ only means that a discrepancy from the hypothesis prediction (e.g., no difference between treatment groups) would be as large *or larger than* that observed no more than 5% of the time if *only* chance were creating the discrepancy (as opposed to a violation of the test hypothesis or a mistaken assumption).

4. *A nonsignificant test result (P > 0.05) means that the test hypothesis is true or should be accepted.*—No! A large *P*-value only suggests that the data are *not* unusual if all the assumptions used to compute the *P*-value (including the test hypothesis) were correct. The same data would also not be unusual under many other hypotheses. Furthermore, even if the test hypothesis is wrong, the *P*-value may be large because it was inflated by a large random error or because of some other erroneous assumption (e.g., the assumption that this *P*-value was not selected for presentation because it was above 0.05). $P > 0.05$ only means that a discrepancy from the hypothesis prediction (e.g., no difference between treatment groups) would be as large *or larger than* that observed more than 5% of the time if *only* chance were

creating the discrepancy.

5. *A large P-value is evidence in favor of the test hypothesis.*—No! In fact, any *P*-value less than 1 implies that the test hypothesis is *not* the hypothesis most compatible with the data, because any other hypothesis with a larger *P*-value would be even more compatible with the data. A *P*-value cannot be said to favor the test hypothesis except in relation to those hypotheses with smaller *P*-values. Furthermore, a large *P*-value often indicates only that the data are incapable of discriminating among many competing hypotheses (as would be seen immediately by examining the range of the confidence interval). For example, many authors will misinterpret $P = 0.70$ from a test of the null hypothesis as evidence for no effect, when in fact it indicates that, even though the null hypothesis is compatible with the data under the assumptions used to compute the *P*-value, it is *not* the hypothesis most compatible with the data—that honor would belong to a hypothesis with $P = 1$. But even if $P = 1$, there will be many other hypotheses that are highly consistent with the data, so that a definitive conclusion of "no association" cannot be deduced from a *P*-value, no matter how large.

6. *A null-hypothesis P-value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated.*—No! Observing $P > 0.05$ for the null hypothesis only means that the null is one among the many hypotheses that have $P > 0.05$. Thus, unless the point estimate (observed association) equals the null value exactly, it is a mistake to conclude from $P > 0.05$ that a study found "no association" or "no evidence" of an effect. If the null *P*-value is less than 1 some association must be present in the data, and one must look at the point estimate to determine the effect size most compatible with the data under the assumed model.

7. *Statistical significance indicates a scientifically or substantively important relation has been detected.*—No! Especially when a study is large, very minor effects or small assumption violations can lead to statistically significant tests of the null hypothesis. Again, a small null *P*-value simply flags the data as being unusual if all the assumptions used to compute it (including the null hypothesis) were correct; but the way the data are unusual might be of no clinical interest. One must look at the confidence interval to determine which effect sizes of scientific or other substantive (e.g., clinical) importance are relatively compatible with the data, given the model.

8. *Lack of statistical significance indicates that the effect size is small.*—No! Especially when a study is small, even large effects may be "drowned in noise" and thus fail to be detected as statistically significant by a statistical test. A large null *P*-value simply flags the data as *not* being unusual if all the assumptions used to compute it (including the test hypothesis) were correct; but the same data will also not be unusual under many other models and hypotheses besides the null. Again, one must look at the confidence interval to determine whether it includes effect sizes of importance.

9. The P-value is the chance of our data occurring if the test hypothesis is true; for example, P = 0.05 means that the observed association would occur only 5% of the time under the test hypothesis.—No! The P-value refers not only to what we observed, but also observations *more extreme* than what we observed (where "extremity" is measured in a particular way). And again, the P-value refers to a data frequency when all the assumptions used to compute it are correct. In addition to the test hypothesis, these assumptions include randomness in sampling, treatment assignment, loss, and missingness, as well as an assumption that the P-value was not selected for presentation based on its size or some other aspect of the results.

10. If you reject the test hypothesis because P ≤ 0.05, the chance you are in error (the chance your "significant finding" is a false positive) is 5%.—No! To see why this description is false, suppose the test hypothesis is in fact true. Then, if you reject it, the chance you are in error is 100%, not 5%. The 5% refers only to how often you would reject it, and therefore be in error, over very many uses of the test across different studies when the test hypothesis and all other assumptions used for the test are true. It does not refer to your single use of the test, which may have been thrown off by assumption violations as well as random errors. This is yet another version of misinterpretation #1.

11. P = 0.05 and P ≤ 0.05 mean the same thing.—No! This is like saying reported height = 2 meters and reported height ≤ 2 meters are the same thing: "height = 2 meters" would include few people and those people would be considered tall, whereas "height ≤ 2 meters" would include most people including small children. Similarly, $P = 0.05$ would be considered a borderline result in terms of statistical significance, whereas $P \leq 0.05$ lumps borderline results together with results very incompatible with the model (e.g., $P = 0.0001$) thus rendering its meaning vague, for no good purpose.

12. P-values are properly reported as inequalities (e.g., report "P < 0.02" when P = 0.015 or report P > 0.05 when P = 0.06 or P = 0.70).—No! This is bad practice because it makes it difficult or impossible for the reader to accurately interpret the statistical result. Only when the P-value is very small (e.g., under 0.001) does an inequality become justifiable: There is little practical difference among very small P-values when the assumptions used to compute P-values are not known with enough certainty to justify such precision, and most methods for computing P-values are not numerically accurate below a certain point.

13. Statistical significance is a property of the phenomenon being studied, and thus statistical tests detect significance.—No! This misinterpretation is promoted when researchers state that they have or have not found "evidence of" a statistically significant effect. The effect being tested either exists or does not exist. "Statistical significance" is a dichotomous description of a P-value (that it is below the chosen cut-off) and thus is a property of a result of a statistical test; it is not a property of the effect or population being studied.

14. One should always use two-sided P-values.—No! Two-sided P-values are designed to test hypotheses that the targeted effect measure equals a specific value (e.g., zero), and is neither above nor below this value. When however the test hypothesis of scientific or practical interest is a one-sided (dividing) hypothesis, a one-sided P-value is appropriate. For example, consider the practical question of whether a new drug is *at least* as good as the standard drug for increasing survival time. This question is one-sided, so testing this hypothesis calls for a one-sided P-value. Nonetheless, because two-sided P-values are the usual default, it will be important to note when and why a one-sided P-value is being used instead.

There are other interpretations of P values that are controversial, in that whether a categorical "No!" is warranted depends on one's philosophy of statistics and the precise meaning given to the terms involved. The disputed claims deserve recognition if one wishes to avoid such controversy.

For example, it has been argued that P-values overstate evidence against test hypotheses, based on directly comparing P-values against certain quantities (likelihood ratios and Bayes factors) that play a central role as evidence measures in Bayesian analysis (Edwards et al. 1963; Berger and Sellke 1987; Edwards 1992; Goodman and Royall 1988; Royall 1997; Sellke et al. 2001; Goodman 1992, 2005; Wagenmakers 2007). Nonetheless, many other statisticians do not accept these quantities as gold standards, and instead point out that P-values summarize crucial evidence needed to gauge the error rates of decisions based on statistical tests (even though they are far from sufficient for making those decisions). Thus, from this frequentist perspective, P-values do not overstate evidence and may even be considered as measuring one aspect of evidence (Cox 1977, 1982; Lehmann 1986; Senn 2001, 2002a; Mayo and Cox 2006), with $1 - P$ measuring evidence against the model used to compute the P-value. See also Murtaugh (2014) and its accompanying discussion.

Common Misinterpretations of P-Value Comparisons and Predictions

Some of the most severe distortions of the scientific literature produced by statistical testing involve erroneous comparison and synthesis of results from different studies or study subgroups. Among the worst are:

15. When the same hypothesis is tested in different studies and none or a minority of the tests are statistically significant (all P > 0.05), the overall evidence supports the hypothesis.—No! This belief is often used to claim that a literature supports no effect when the opposite is case. It reflects a tendency of researchers to "overestimate the power of most research" (Hedges and Olkin 1980). In reality, every study could fail to reach statistical significance and yet when combined show a statistically significant association and persuasive evidence of an effect. For example, if there were five studies each with $P = 0.10$, none would be significant at 0.05 level; but when these P-values are combined using the Fisher formula (Cox and Hinkley 1974, p. 80), the overall P-value would be 0.01. There are many real ex-

amples of persuasive evidence for important effects when few studies or even no study reported "statistically significant" associations (e.g., Chalmers and Lau 1996; Maheshwari et al. 2007). Thus, lack of statistical significance of individual studies should not be taken as implying that the totality of evidence supports no effect.

16. When the same hypothesis is tested in two different populations and the resulting P-values are on opposite sides of 0.05, the results are conflicting.—No! Statistical tests are sensitive to many differences between study populations that are irrelevant to whether their results are in agreement, such as the sizes of compared groups in each population. As a consequence, two studies may provide very different P-values for the same test hypothesis and yet be in perfect agreement (e.g., may show identical observed associations). For example, suppose we had two randomized trials A and B of a treatment, identical except that trial A had a known standard error of 2 for the mean difference between treatment groups whereas trial B had a known standard error of 1 for the difference. If both trials observed a difference between treatment groups of exactly 3, the usual normal test would produce $P = 0.13$ in A but $P = 0.003$ in B. Despite their difference in P-values, the test of the hypothesis of no difference in effect across studies would have $P = 1$, reflecting the perfect agreement of the observed mean differences from the studies. Differences between results must be evaluated by directly, for example by estimating and testing those differences to produce a confidence interval and a P-value comparing the results (often called analysis of heterogeneity, interaction, or modification).

17. When the same hypothesis is tested in two different populations and the same P-values are obtained, the results are in agreement.—No! Again, tests are sensitive to many differences between populations that are irrelevant to whether their results are in agreement. Two different studies may even exhibit identical P-values for testing the same hypothesis yet also exhibit clearly different observed associations. For example, suppose randomized experiment A observed a mean difference between treatment groups of 3.00 with standard error 1.00, while B observed a mean difference of 12.00 with standard error 4.00. Then the standard normal test would produce $P = 0.003$ in both; yet the test of the hypothesis of no difference in effect across studies gives $P = 0.03$, reflecting the large difference $(12.00 - 3.00 = 9.00)$ between the mean differences.

18. If one observes a small P-value, there is a good chance that the next study will produce a P-value at least as small for the same hypothesis.—No! This is false even under the ideal condition that both studies are independent and all assumptions including the test hypothesis are correct in both studies. In that case, if (say) one observes $P = 0.03$, the chance that the new study will show $P \leq 0.03$ is only 3%; thus the chance the new study will show a P-value as small or smaller (the "replication probability") is exactly the observed P-value! If on the other hand the small P-value arose solely because the true effect exactly equaled its observed estimate, there would be a 50% chance that a repeat experiment of identical design would have

a larger P-value (Goodman 1992). In general, the size of the new P-value will be extremely sensitive to the study size and the extent to which the test hypothesis or other assumptions are violated in the new study (Senn 2002a); in particular, P may be very small or very large depending on whether the study and the violations are large or small.

Finally, although it is (we hope obviously) wrong to do so, one sometimes sees the null hypothesis compared with another (alternative) hypothesis using a two-sided P-value for the null and a one-sided P-value for the alternative. This comparison is biased in favor of the null in that the two-sided test will falsely reject the null only half as often as the one-sided test will falsely reject the alternative (again, under all the assumptions used for testing).

Common Misinterpretations of Confidence Intervals

Most of the above misinterpretations translate into an analogous misinterpretation for confidence intervals. For example, another misinterpretation of $P > 0.05$ is that it means the test hypothesis has only a 5% chance of being false, which in terms of a confidence interval becomes the common fallacy:

19. The specific 95% confidence interval presented by a study has a 95% chance of containing the true effect size.—No! A reported confidence interval is a range between two numbers. The frequency with which an observed interval (e.g., 0.72 to 2.88) contains the true effect is either 100% if the true effect is within the interval or 0% if not; the 95% refers only to how often 95% confidence intervals computed from very many studies would contain the true size *if all the assumptions used to compute the intervals were correct*. It is possible to compute an interval that can be interpreted as having 95% probability of containing the true value; nonetheless, such computations require not only the assumptions used to compute the confidence interval, but also further assumptions about the size of effects in the model. These further assumptions are summarized in what is called a *prior distribution*, and the resulting intervals are usually called *Bayesian posterior (or credible) intervals* to distinguish them from confidence intervals (e.g., see Rothman et al. 2008, Ch. 13 and 18).

Symmetrically, the misinterpretation of a small P-value as disproving the test hypothesis could be translated into:

20. An effect size outside the 95% confidence interval has been refuted (or excluded) by the data.—No! As with the P-value, the confidence interval is computed from many assumptions, the violation of which may have led to the results. Thus it is the combination of the data with the assumptions, along with the arbitrary 95% criterion, that are needed to declare an effect size outside the interval is in some way incompatible with the observations. Even then, judgements as extreme as saying the effect size has been refuted or excluded will require even stronger conditions.

As with P-values, nave comparison of confidence intervals

can be highly misleading:

21. If two confidence intervals overlap, the difference between two estimates or studies is not significant.—No! The 95% confidence intervals from two subgroups or studies may overlap substantially and yet the test for difference between them may still produce $P < 0.05$. Suppose for example, two 95% confidence intervals for means from normal populations with known variances are (1.04, 4.96) and (4.16, 19.84); these intervals overlap, yet the test of the hypothesis of no difference in effect across studies gives $P = 0.03$. As with P-values, comparison between groups requires statistics that directly test and estimate the differences across groups. It can, however, be noted that if the two 95% confidence intervals fail to overlap, then when using the same assumptions used to compute the confidence intervals we will find $P < 0.05$ for the difference; and if one of the 95% intervals contains the point estimate from the other group or study, we will find $P > 0.05$ for the difference.

Finally, as with P-values, the replication properties of confidence intervals are usually misunderstood:

22. An observed 95% confidence interval predicts that 95% of the estimates from future studies will fall inside the observed interval.—No! This statement is wrong in several ways. Most importantly, under the model, 95% is the frequency with which *other unobserved* intervals will contain the *true effect*, not how frequently the one interval being presented will contain future estimates. In fact, even under ideal conditions the chance that a future estimate will fall within the current interval will usually be much less than 95%. For example, if two independent studies of the same quantity provide unbiased normal point estimates with the same standard errors, the chance that the 95% confidence interval for the first study contains the point estimate from the second is 83% (which is the chance that the difference between the two estimates is less than 1.96 standard errors). Again, an observed interval either does or does not contain the true effect; the 95% refers only to how often 95% confidence intervals computed from very many studies would contain the true effect if all the assumptions used to compute the intervals were correct.

23. If one 95% confidence interval includes the null value and another excludes that value, the interval excluding the null is the more precise one.—No! When the model is correct, precision of statistical estimation is measured directly by confidence interval *width* (measured on the appropriate scale). It is not a matter of inclusion or exclusion of the null or any other value. Consider two 95% confidence intervals for a difference in means, one with limits of 5 and 40, the other with limits of −5 and 10. The first interval excludes the null value of 0, but is 30 units wide. The second includes the null value, but is half as wide and therefore much more precise.

In addition to the above misinterpretations, 95% confidence intervals force the 0.05-level cutoff on the reader, lumping together all effect sizes with $P > 0.05$, and in this way are as bad as presenting P-values as dichotomies. Nonetheless, many authors agree that confidence intervals are superior to tests and P-values because they allow one to shift focus away from the null hypothesis, toward the full range of effect sizes compatible with the data—a shift recommended by many authors and a growing number of journals. Another way to bring attention to nonnull hypotheses is to present their P-values; for example, one could provide or demand P-values for those effect sizes that are recognized as scientifically reasonable alternatives to the null.

As with P-values, further cautions are needed to avoid misinterpreting confidence intervals as providing sharp answers when none are warranted. The hypothesis which says the point estimate is the correct effect will have the largest P-value ($P = 1$ in most cases), and hypotheses inside a confidence interval will have higher P-values than hypotheses outside the interval. The P-values will vary greatly, however, among hypotheses inside the interval, as well as among hypotheses on the outside. Also, two hypotheses may have nearly equal P-values even though one of the hypotheses is inside the interval and the other is outside. Thus, if we use P-values to measure compatibility of hypotheses with data and wish to compare hypotheses with this measure, we need to examine their P-values directly, not simply ask whether the hypotheses are inside or outside the interval. This need is particularly acute when (as usual) one of the hypotheses under scrutiny is a null hypothesis.

Common Misinterpretations of Power

The *power* of a test to detect a correct alternative hypothesis is the pre-study probability that the test will reject the test hypothesis (e.g., the probability that P will not exceed a prespecified cut-off such as 0.05). (The corresponding prestudy probability of failing to reject the test hypothesis when the alternative is correct is one minus the power, also known as the Type-II or beta error rate; see Lehmann 1986.) As with P-values and confidence intervals, this probability is defined over repetitions of the same study design and so is a frequency probability. One source of reasonable alternative hypotheses are the effect sizes that were used to compute power in the study proposal. Prestudy power calculations do not, however, measure the compatibility of these alternatives with the data actually observed, while power calculated from the observed data is a direct (if obscure) transformation of the null P-value and so provides no test of the alternatives. Thus, presentation of power does not obviate the need to provide interval estimates and direct tests of the alternatives.

For these reasons, many authors have condemned use of power to interpret estimates and statistical tests (e.g., Cox 1958; Smith and Bates 1992; Goodman 1994; Goodman and Berlin 1994; Hoenig and Heisey 2001; Senn 2002b; Greenland 2012a), arguing that (in contrast to confidence intervals) it distracts attention from direct comparisons of hypotheses and introduces new misinterpretations, such as:

24. If you accept the null hypothesis because the null P-value exceeds 0.05 and the power of your test is 90%, the chance you are in error (the chance that your finding is a false negative) is 10%.—No! If the null hypothesis is false and you accept

it, the chance you are in error is 100%, not 10%. Conversely, if the null hypothesis is true and you accept it, the chance you are in error is 0%. The 10% refers only to how often you would be in error over very many uses of the test across different studies when the particular alternative used to compute power is correct *and* all other assumptions used for the test are correct in all the studies. It does not refer to your single use of the test or your error rate under any alternative effect size other than the one used to compute power.

It can be especially misleading to compare results for two hypotheses by presenting a test or *P*-value for one and power for the other. For example, testing the null by seeing whether $P \leq 0.05$ with a power less than $1-0.05 = 0.95$ for the alternative (as done routinely) will bias the comparison in favor of the null because it entails a lower probability of incorrectly rejecting the null (0.05) than of incorrectly accepting the null when the alternative is correct. Thus, claims about relative support or evidence need to be based on direct and comparable measures of support or evidence for both hypotheses, otherwise mistakes like the following will occur:

25. If the null P-value exceeds 0.05 and the power of this test is 90% at an alternative, the results support the null over the alternative.—This claim seems intuitive to many, but counterexamples are easy to construct in which the null *P*-value is between 0.05 and 0.10, and yet there are alternatives whose own *P*-value exceeds 0.10 and for which the power is 0.90. Parallel results ensue for other accepted measures of compatibility, evidence, and support, indicating that the data show lower compatibility with and more evidence against the null than the alternative, despite the fact that the null *P*-value is "not significant" at the 0.05 alpha level and the power against the alternative is "very high" (Greenland, 2012a).

Despite its shortcomings for interpreting current data, power can be useful for designing studies and for understanding why replication of "statistical significance" will often fail even under ideal conditions. Studies are often designed or claimed to have 80% power against a key alternative when using a 0.05 significance level, although in execution often have less power due to unanticipated problems such as low subject recruitment. Thus, if the alternative is correct and the actual power of two studies is 80%, the chance that the studies will both show $P \leq 0.05$ will at best be only $0.80(0.80) = 64\%$; furthermore, the chance that one study shows $P \leq 0.05$ and the other does not (and thus will be misinterpreted as showing conflicting results) is $2(0.80)0.20 = 32\%$ or about 1 chance in 3. Similar calculations taking account of typical problems suggest that one could anticipate a "replication crisis" even if there were no publication or reporting bias, simply because current design and testing conventions treat individual study results as dichotomous outputs of "significant"/"nonsignificant" or "reject"/"accept."

A Statistical Model is Much More Than an Equation with Greek Letters

The above list could be expanded by reviewing the research literature. We will however turn to direct discussion of an issue that has been receiving more attention of late, yet is still widely overlooked or interpreted too narrowly in statistical teaching and presentations: That the statistical model used to obtain the results is correct.

Too often, the full statistical model is treated as a simple regression or structural equation in which effects are represented by parameters denoted by Greek letters. "Model checking" is then limited to tests of fit or testing additional terms for the model. Yet these tests of fit themselves make further assumptions that should be seen as part of the full model. For example, all common tests and confidence intervals depend on assumptions of random selection for observation or treatment and random loss or missingness within levels of controlled covariates. These assumptions have gradually come under scrutiny via sensitivity and bias analysis (e.g., Lash et al. 2014), but such methods remain far removed from the basic statistical training given to most researchers.

Less often stated is the even more crucial assumption that the analyses themselves were not guided toward finding nonsignificance or significance (analysis bias), and that the analysis results were not reported based on their nonsignificance or significance (reporting bias and publication bias). Selective reporting renders false even the limited ideal meanings of statistical significance, *P*-values, and confidence intervals. Because author decisions to report and editorial decisions to publish results often depend on whether the *P*-value is above or below 0.05, selective reporting has been identified as a major problem in large segments of the scientific literature (Dwan et al. 2013; Page et al. 2014; You et al. 2012).

Although this selection problem has also been subject to sensitivity analysis, there has been a bias in studies of reporting and publication bias: It is usually assumed that these biases favor significance. This assumption is of course correct when (as is often the case) researchers select results for presentation when $P \leq 0.05$, a practice that tends to exaggerate associations (Button et al. 2013; Eyding et al. 2010; Land 1980; Land 1981). Nonetheless, bias in favor of reporting $P \leq 0.05$ is not always plausible let alone supported by evidence or common sense. For example, one might expect selection for $P > 0.05$ in publications funded by those with stakes in acceptance of the null hypothesis (a practice which tends to understate associations); in accord with that expectation, some empirical studies have observed smaller estimates and "nonsignificance" more often in such publications than in other studies (Eyding et al. 2010; Greenland 2009; Xu et al. 2013).

Addressing such problems would require far more political will and effort than addressing misinterpretation of statistics, such as enforcing registration of trials, along with open data and analysis code from all completed studies (as in the AllTrials initiative, *http://www.alltrials.net/*). In the meantime, readers are advised to consider the entire context in which research reports are produced and appear when interpreting the statistics and conclusions offered by the reports.

Conclusions

Upon realizing that statistical tests are usually misinterpreted, one may wonder what if anything these tests do for science. They were originally intended to account for random variability as a source of error, thereby sounding a note of caution against overinterpretation of observed associations as true effects or as stronger evidence against null hypotheses than was warranted. But before long that use was turned on its head to provide fallacious support for null hypotheses in the form of "failure to achieve" or "failure to attain" statistical significance.

We have no doubt that the founders of modern statistical testing would be horrified by common treatments of their invention. In their first paper describing their binary approach to statistical testing, Neyman and Pearson (1928) wrote that "it is doubtful whether the knowledge that [a P-value] was really 0.03 (or 0.06), rather than 0.05 … would in fact ever modify our judgment" and that "The tests themselves give no final verdict, but as tools help the worker who is using them to form his final decision." Pearson (1955) later added, "No doubt we could more aptly have said, 'his final or provisional decision'." Fisher (1956, p. 42) went further, saying "No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas." Yet fallacious and ritualistic use of tests continued to spread, including beliefs that whether P was above or below 0.05 was a universal arbiter of discovery. Thus by 1965, Hill (1965) lamented that "too often we weaken our capacity to interpret data and to take reasonable decisions whatever the value of P. And far too often we deduce 'no difference' from 'no significant difference'."

In response, it has been argued that some misinterpretations are harmless in tightly controlled experiments on well-understood systems, where the test hypothesis may have special support from established theories (e.g., Mendelian genetics) and in which every other assumption (such as random allocation) is forced to hold by careful design and execution of the study. But it has long been asserted that the harms of statistical testing in more uncontrollable and amorphous research settings (such as social-science, health, and medical fields) have far outweighed its benefits, leading to calls for banning such tests in research reports—again, with one journal banning confidence intervals as well as P-values (Trafimow and Marks 2015).

Given, however, the deep entrenchment of statistical testing, as well as the absence of generally accepted alternative methods, there have been many attempts to salvage P-values by detaching them from their use in significance tests. One approach is to focus on P-values as continuous measures of compatibility, as described earlier. Although this approach has its own limitations (as described in points 1, 2, 5, 9, 17, and 18), it avoids misconceptions arising from comparison of P-values with arbitrary cutoffs such as 0.05 (as described in points 3, 4, 6–8, 10–13, 15, 16, 21, and 23–25). Another approach is to teach and use correct relations of P-values to hypothesis probabilities. For example, under common statistical models, one-sided P-values can provide lower bounds on probabilities for hypotheses about effect directions (Casella and Berger 1987ab; Greenland and

Poole 2013ab). Whether such reinterpretations can eventually replace common misinterpretations to good effect remains to be seen.

A shift in emphasis from hypothesis testing to estimation has been promoted as a simple and relatively safe way to improve practice (Yates 1951; Rothman 1978; Altman et al. 2000; Poole 2001; Cumming 2011), resulting in increasing use of confidence intervals and editorial demands for them; nonetheless, this shift has brought to the fore misinterpretations of intervals such as 19–23 above (Morey et al. 2015). Other approaches combine tests of the null with further calculations involving both null and alternative hypotheses (Rosenthal and Rubin 1994; Mayo and Spanos 2006); such calculations may, however, may bring with them further misinterpretations similar to those described above for power, as well as greater complexity.

Meanwhile, in the hopes of minimizing harms of current practice, we can offer several guidelines for users and readers of statistics, and re-emphasize some key warnings from our list of misinterpretations:

a) Correct and careful interpretation of statistical tests demands examining the sizes of effect estimates and confidence limits, as well as precise P-values (not just whether P-values are above or below 0.05 or some other threshold).

b) Careful interpretation also demands critical examination of the assumptions and conventions used for the statistical analysis—not just the usual statistical assumptions, but also the hidden assumptions about how results were generated and chosen for presentation.

c) It is simply false to claim that statistically nonsignificant results support a test hypothesis, because the same results may be even more compatible with alternative hypotheses—even if the power of the test is high for those alternatives.

d) Interval estimates aid in evaluating whether the data are capable of discriminating among various hypotheses about effect sizes, or whether statistical results have been misrepresented as supporting one hypothesis when those results are better explained by other hypotheses (see points 4–6). We caution however that confidence intervals are often only a first step in these tasks. To compare hypotheses in light of the data and the statistical model it may be necessary to calculate the P-value (or relative likelihood) of each hypothesis. We further caution that confidence intervals provide only a best-case measure of the uncertainty or ambiguity left by the data, insofar as they depend on an uncertain statistical model.

e) Correct statistical evaluation of multiple studies requires a pooled analysis or meta-analysis that deals correctly with study biases (Whitehead 2002; Borenstein et al. 2009; Chen and Peace 2013; Cooper et al. 2009; Greenland and O'Rourke 2008; Petitti 2000; Schmidt and Hunter 2014; Sterne 2009). Even when this is done, however, all the earlier cautions apply. Furthermore, the outcome of any statistical procedure is but one of many considerations that must be evaluated when examining the totality of evidence. In

particular, statistical significance is neither necessary nor sufficient for determining the scientific or practical significance of a set of observations. This view was affirmed unanimously by the U.S. Supreme Court, (Matrixx Initiatives, Inc., et al. v. Siracusano et al. No. 091156. Argued January 10, 2011, Decided March 22, 2011), and can be seen in our earlier quotes from Neyman and Pearson.

f) Any opinion offered about the *probability, likelihood, certainty*, or similar property for a hypothesis *cannot* be derived from statistical methods alone. In particular, significance tests and confidence intervals do not by themselves provide a logically sound basis for concluding an effect is present or absent with certainty or a given probability. This point should be borne in mind whenever one sees a conclusion framed as a statement of probability, likelihood, or certainty about a hypothesis. Information about the hypothesis beyond that contained in the analyzed data and in conventional statistical models (which give only data probabilities) must be used to reach such a conclusion; that information should be explicitly acknowledged and described by those offering the conclusion. Bayesian statistics offers methods that attempt to incorporate the needed information directly into the statistical model; they have not however achieved the popularity of P-values and confidence intervals, in part because of philosophical objections and in part because no conventions have become established for their use.

g) All statistical methods (whether frequentist or Bayesian, or for testing or estimation, or for inference or decision) make extensive assumptions about the sequence of events that led to the results presented—not only in the data generation, but in the analysis choices. Thus, to allow critical evaluation, research reports (including meta-analyses) should describe in detail the full sequence of events that led to the statistics presented, including the motivation for the study, its design, the original analysis plan, the criteria used to include and exclude subjects (or studies) and data, and a thorough description of all the analyses that were conducted.

In closing, we note that no statistical method is immune to misinterpretation and misuse, but prudent users of statistics will avoid approaches especially prone to serious abuse. In this regard, we join others in singling out the degradation of P-values into "significant" and "nonsignificant" as an especially pernicious statistical practice (Weinberg 2001).

References

Altman, D.G., and Bland, J.M. (1995), "Absence of Evidence is not Evidence of Absence," *British Medical Journal*, 311, 485.

Altman, D.G., Machin, D., Bryant, T.N., and Gardner, M.J. (eds.) (2000), *Statistics with Confidence* (2nd ed.), London: BMJ Books.

Anscombe, F.J. (1990), "The Summarizing of Clinical Experiments by Significance Levels," *Statistics in Medicine*, 9, 703–708.

Ashworth, A. (2015), "Veto on the Use of Null Hypothesis Testing and *p* Intervals: Right or Wrong?" Taylor & Francis Editor Resources online, *http://editorresources.taylorandfrancisgroup.com/veto-on-the-use-of-null-hypothesis-testing-and-p-intervals-right-or-wrong/*, accessed Feb. 27, 2016.

Atkins, L., and Jarrett, D. (1979), "The Significance of 'Significance Tests,'" in *Demystifying Social Statistics*, Irvine, J., Miles, I., and Evans, J., eds. London: Pluto Press.

Bakan, D. (1966), "The Test of Significance in Psychological Research," *Psychological Bulletin*, 66, 423–437.

Bandt, C.L., and Boen, J.R. (1972), "A Prevalent Misconception About Sample Size, Statistical Significance, and Clinical Importance," *Journal of Periodontal Research*, 43, 181–183.

Berger, J.O., and Sellke, T.M. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of *P*-values and Evidence," *Journal of the American Statistical Association*, 82, 112–139.

Berkson, J. (1942), "Tests of Significance Considered as Evidence," *Journal of the American Statistical Association*, 37, 325–335.

Bland, J.M., and Altman, D.G. (2015), "Best (but oft forgotten) Practices: Testing for Treatment Effects in Randomized Trials by Separate Analyses of Changes from Baseline in Each Group is a Misleading Approach," *American Journal of Clinical Nutrition*, 102, 991–994.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., and Rothstein, H.R. (2009), *Introduction to Meta-Analysis*, New York: Wiley.

Button, K., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafó, M.R. (2013), "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience," *Nature Reviews Neuroscience*, 14, 365–376.

Casella, G., and Berger, R.L. (1987a), "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem," *Journal of the American Statistical Association*, 82, 106–111.

Casella, G., and Berger, R.L. (1987b), "Comment," *Stat Sci*, 2, 344–417.

Chalmers, T.C., and Lau, J. (1996), "Changes in Clinical Trials Mandated by the Advent of Meta-analysis," *Statistics in Medicine*, 15, 1263–1268.

Chen, D.-G., and Peace, K.E. (2013), *Applied Meta-Analysis with R*, New York: Chapman & Hall/CRC.

Chia, K.S. (1997), "'Significant-itis' An Obsession with the *P*-Value," *Scandinavian Journal of Work, Environment & Health*, 23, 152–154.

Cooper, H., Hedges, L.V., and Valentine, J.C. (2009), *The Handbook of Research Synthesis and Meta-Analysis*, Thousand Oaks, CA: Sage.

Cohen, J. (1994), "The Earth is Round ($p < 0.05$)," *American Psychology*, 47, 997–1003.

Cornfield, J. (1966), "Sequential Trials, Sequential Analysis, and the Likelihood Principle," *The American Statistician*, 25, 617–657.

Cox, D.R. (1958), *The Planning of Experiments*, New York: Wiley, p. 161.

——— (1977), "The Role of Significance Tests" (with discussion), *Scandinavian Journal of Statistics*, 4, 49–70.

——— (1982), "Statistical Significance Tests," *British Journal of Clinical Pharmacology*, 14, 325–331.

Cox, D.R., and Hinkley, D.V. (1974), *Theoretical Statistics*, New York: Chapman and Hall.

Cumming, G. (2011), *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*, London: Routledge.

Dickersin, K. (1990), "The Existence of Publication Bias and Risk Factors for its Occurrence," *Journal of the American Medical Association*, 263, 1385–1389.

Dwan, K., Gamble, C., Williamson, P.R., Kirkham, J.J.; Reporting Bias Group (2013), "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias—An Updated Review," *PLoS One*, 8:e66844.

Edwards, A.W.F. (1992), *Likelihood* (2nd ed.), Baltimore: Johns Hopkins University Press.

Edwards, W., Lindman, H., and Savage, L.J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242.

Evans, S.J.W., Mills, P., and Dawson, J. (1988), "The End of the *P*-value?" *British Heart Journal*, 60,177–180.

Eyding, D., Lelgemann, M., Grouven, U., Härter, M., Kromp, M., Kaiser, T., Kerekes, M.F., Gerken, M., and Wieseler, B. (2010), "Reboxetine for Acute Treatment of Major Depression: Systematic Review and Meta-analysis of Published and Unpublished Placebo and Selective Serotonin Reuptake In-

hibitor Controlled Trials," *British Medical Journal*, 341, c4737.

Fidler, F., and Loftus, G.R. (2009), "Why Figures with Error Bars Should Replace *p* Values: Some Conceptual Arguments and Empirical Demonstrations," *Journal of Psychology*, 217, 27–37.

Fisher, R. A. (1956), *Statistical Methods and Scientific Inference*, Edinburgh, UK: Oliver & Boyd.

Flanagan, O. (2015), "Journal's Ban on Null Hypothesis Significance Testing: Reactions from the Statistical Arena," *Stats Life* online, accessed 27 Feb. 2016.

Freedman, D.A., Pisani, R., and Purves, R. (2007), *Statistics* (4th ed.), New York: Norton.

Gardner, M.A., and Altman, D.G. (1986), "Confidence Intervals Rather than *P* Values: Estimation Rather than Hypothesis Testing," *British Medical Journal*, 292, 746–750.

Gelman, A. (2013), "*P*-Values and Statistical Practice," *Epidemiology*, 24, 69–72.

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science: Data-Dependent Analysis—A 'Garden of Forking Paths'—Explains why Many Statistically Significant Comparisons Don't Hold Up," *American Scientist*, 102, 460–465. Erratum at *http://andrewgelman.com/2014/10/14/didnt-say-part-2/*, accessed Feb. 27, 2016.

Gelman, A., and Stern, H.S. (2006), "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant," *The American Statistician*, 60, 328–331.

Gibbons, J.D., and Pratt, J.W. (1975), "P-Values: Interpretation and Methodology," *The American Statistician*, 29, 20–25.

Gigerenzer, G. (2004), "Mindless Statistics," *Journal of Socioeconomics*, 33, 567–606.

Gigerenzer, G., and Marewski, J.N. (2015), "Surrogate Science: The Idol of a Universal Method for Scientific Inference," *Journal of Management*, 41, 421–440.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., and Kruger, L. (1990), *The Empire of Chance: How Probability Changed Science and Everyday Life*, New York: Cambridge University Press.

Goodman, S.N. (1992), "A Comment on Replication, *p*-values and Evidence," *Statistics in Medicine*, 11, 875–879.

——— (1993), "P-values, Hypothesis Tests and Likelihood: Implications for Epidemiology of a Neglected Historical Debate," *American Journal of Epidemiology*, 137, 485–496.

——— (1994), "Letter to the editor re Smith and Bates, *Epidemiology*, 5, 266–268.

——— (1999), "Towards Evidence-Based Medical Statistics, I: The *P*-value Fallacy," *Annals of Internal Medicine*, 130, 995–1004.

——— (2005), "Introduction to Bayesian Methods I: Measuring the Strength of Evidence," *Clinical Trials*, 2, 282–290.

——— (2008), "A Dirty Dozen: Twelve *P*-value Misconceptions," *Seminars in Hematology*, 45, 135–140.

Goodman, S.N., and Berlin, J. (1994), "The Use of Predicted Confidence Intervals when Planning Experiments and the Misuse of Power when Interpreting Results," *Annals of Internal Medicine*, 121, 200–206.

Goodman, S.N., and Royall, R. (1988), "Evidence and Scientific Research," *American Journal of Public Health*, 78, 1568–1574.

Greenland, S. (2009), "Dealing with Uncertainty About iInvestigator Bias: Disclosure is Informative," *Journal of Epidemiology and Community Health*, 63, 593–598.

——— (2011), "Null Misinterpretation in Statistical Testing and its Impact on Health Risk Assessment," *Preventive Medicine*, 53, 225–228.

——— (2012a), "Nonsignificance Plus High Power Does not Imply Support for the Null over the Alternative," *Annals of Epidemiology*, 22, 364–368.

——— (2012b), "Transparency and Disclosure, Neutrality and Balance: Shared Values or Just Shared Words?" *Journal of Epidemiology and Community Health*, 66, 967–970.

Greenland, S., and O'Rourke, K. (2008), "Meta-analysis," in *Modern Epidemiology* (3rd ed.), Rothman, K.J., Greenland, S., Lash, T.L., eds., Philadelphia: Lippincott-Wolters-Kluwer, pp. 682–685.

Greenland, S., and Poole, C. (2011), "Problems in Common Interpretations of Statistics in Scientific Articles, Expert Reports, and Testimony," *Jurimetrics*, 51, 113–129.

——— (2013a), "Living with *P*-values: Resurrecting a Bayesian Perspective on Frequentist Statistics," *Epidemiology*, 24, 62–68.

——— (2013b), "Living with Statistics in Observational Research," *Epidemiology*, 24, 73–78.

Grieve, A.P. (2015), "How to Test Hypotheses if You Must," *Pharmaceutical Statistics*, 14, 139–150.

Hanley, J.A. (1994), Letter to the Editor re Smith and Bates," *Epidemiology*, 5, 264–266.

Harlow, L.L., Mulaik, S.A., and Steiger, J.H. (1997), "What if There Were No Significance Tests?" *Psychology Press*.

Hauer, E. (2003), "The Harm Done by Tests of Significance," *Accident Analysis & Prevention*, 36, 495–500.

Hedges, L.V., and Olkin, I. (1980), "Vote-Counting Methods in Research Synthesis," *Psychological Bulletin*, 88, 359–369.

Hill, A.B. (1965), "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine*, 58, 295–300.

Hoekstra, R., Finch, S., Kiers, H.A.L., and Johnson, A. (2006), "Probability as Certainty: Dichotomous Thinking and the Misuse of *p*-Values," *Psychological Bulletin Review*, 13, 1033–1037.

Hoenig, J.M., and Heisey, D.M. (2001), "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis," *The American Statistician*, 55, 19–24.

Hogben, L. (1957), *Statistical Theory*, London: Allen and Unwin.

Hurlbert, S.H., and Lombardi, C.M. (2009), "Final Collapse of the Neyman-Pearson Decision Theoretic Framework and Rise of the neoFisherian," *Annales Zoologici Fennici*, 46, 311–349.

Kaye, D.H. (1986), "Is Proof of Statistical Significance Relevant?" *Washington Law Review*, 61, 1333–1366.

Kaye, D.H., and Freedman, D.A. (2011), "Reference Guide on Statistics," in *Reference Manual on Scientific Evidence* (3rd ed.), Washington, DC: Federal Judicial Center, 211–302.

Kline, R.B. (2013), *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences*, Washington, DC: American Psychological Association.

Lambdin, C. (2012), "Significance Tests as Sorcery: Science is Empirical—Significance Tests are Not," *Theory & Psychology*, 22, 67–90.

Land, C.E. (1980), "Estimating Cancer Risks from Low Doses of Ionizing Radiation," *Science*, 209, 1197–1203.

——— (1981), "Statistical Limitations in Relation to Sample Size," *Environmental Health Perspectives*, 42, 15–21.

Lang, J.M., Rothman, K.J., and Cann, C.I. (1998), "That Confounded *P*-Value," *Epidemiology*, 9, 7–8.

Langman, M.J.S. (1986), "Towards Estimation and Confidence Intervals," *BMJ*, 292, 716.

Lash, T.L., Fox, M.P., Maclehose, R.F., Maldonado, G., McCandless, L.C., and Greenland, S. (2014), "Good Practices for Quantitative Bias Analysis," *International Journal of Epidemiology*, 43, 1969–1985.

Lecoutre, M.-P., Poitevineau, J., and Lecoutre, B. (2003), "Even Statisticians are not Immune to Misinterpretations of Null Hypothesis Tests," *International Journal of Psychology*, 38, 37–45.

Lehmann, E.L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York, Wiley.

Lew, M.J. (2012), "Bad Statistical Practice in Pharmacology (and Other Basic Biomedical Disciplines): You Probably Don't Know *P*," *British Journal of Pharmacology*, 166, 1559–1567.

Loftus, G.R. (1996), "Psychology Will be a Much Better Science When We Change the Way We Analyze Data," *Current Directions in Psychology*, 5, 161–171.

Maheshwari, S., Sarraj, A., Kramer, J., and El-Serag, H.B. (2007), "Oral Contraception and the Risk of Hepatocellular Carcinoma," *Journal Hepatology*, 47, 506–513.

Marshall, S.W. (2006), "Commentary on Making Meaningful Inferences About Magnitudes," *Sportscience*, 9, 43–44.

Matthews, J.N.S., and Altman, D.G. (1996a), "Interaction 2: Compare Effect Sizes not *P* Values," *British Medical Journal*, 313, 808.

——— (1996b), "Interaction 3: How to Examine Heterogeneity," *British Medical Journal*, 313, 862.

Mayo, D.G., and Cox, D.R. (2006), "Frequentist Statistics as a Theory of Inductive Inference," in *Optimality: The Second Erich L. Lehmann Symposium*, Lecture Notes-Monograph Series, J. Rojo (ed.), Hayward, CA: Institute of Mathematical Statistics (IMS) 49, 77–97.

Mayo, D.G., and Spanos, A. (2006), "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction," *British Journal of Philosophical Science*, 57, 323–357.

Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., and Wagenmakers, E.-J. (in press), "The Fallacy of Placing Confidence in Confidence Intervals," *Psychological Bulletin Review*.

Morrison, D.E., and Henkel, R.E. (eds.) (1970), *The Significance Test Controversy* , Chicago: Aldine.

Murtaugh, P.A. (2014), "In Defense of *P*-Values" (with discussion), *Ecology*, 95, 611–653.

Neyman, J. (1937), "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability," *Philosophical Transactions of the Royal Society of London A*, 236, 333–380.

Neyman, J., and Pearson, E.S. (1928), "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I," *Biometrika*, 20A, 175–240.

Oakes, M. (1986), *Statistical Inference: A Commentary for the Social and Behavioural Sciences*, Chichester: Wiley.

Page, M.J., McKenzie, J.E., Kirkham, J., Dwan, K., Kramer, S., Green, S., and Forbes, A. (2014), "Bias Due to Selective Inclusion and Reporting of Outcomes and Analyses in Systematic Reviews of Randomised Trials of Healthcare Interventions," *Cochrane Database System Reviews*, 10:MR000035.

Peace, K. (1988), "Some Thoughts on One-Tailed Tests," *Biometrics*, 44, 911–912.

Pearson, E.S. (1955), "Statistical Concepts in the Relation to Reality," *Journal fo the Royal Statistical Society*, Series B, 17, 204–207.

Petitti, D.B. (2000), *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine* (2nd ed.), New York: Oxford University Press.

Phillips, C.V. (2004), "Publication Bias In Situ," *BMC Medical Research and Methodology*, 4, 20.

Pocock, S.J., Hughes, M.D., and Lee, R.J. (1987), "Statistical Problems in the Reporting of Clinical Trials," *New England Journal of Medicine*, 317, 426–432.

Pocock, S.J., and Ware, J.H. (2009), "Translating Statistical Findings into Plain English," *The Lancet*, 373, 1926–1928.

Poole, C. (1987a), "Beyond the Confidence Interval," *American Journal of Public Health*, 77, 195–199.

——— (1987b), "Confidence Intervals Exclude Nothing," *American Journal of Public Health*, 77:, 492–493.

——— (2001), "Low *P*-Values or Narrow Confidence Intervals: Which are More Durable?", *Epidemiology*, 12, 291–294.

Pratt, J.W. (1965), "Bayesian Interpretation of Standard Inference Statements," *Journal of the Royal Statistical Society*, Series B, 27, 169–203.

Rosenthal, R., and Rubin, D.B. (1994), "The Counternull Value of an Effect Size: A New Statistic," *Psychological Science*, 5, 329–334.

Rosnow, R.L., and Rosenthal, R. (1989), "Statistical Procedures and the Justification of Knowledge in Psychological Science," *American Psychologist*, 44, 1276–1284.

Rothman, K.J. (1978), "A Show of Confidence," *New England School of Medicine*, 299, 1362–1363.

——— (1986), "Significance Questing," *Annals of Internal Medicine*, 105, 445–447.

Rothman, K.J., Greenland, S., and Lash, T.L. (2008), *Modern Epidemiology* (3rd ed.), Philadelphia, PA: Lippincott-Wolters-Kluwer.

Royall, R. (1997), *Statistical Evidence,* New York: Chapman and Hall.

Rozeboom, W.M. (1960), "The Fallacy of Null-Hypothesis Significance Test," *Psychological Bulletin*, 57, 416–428.

Salsburg, D.S. (1985), "The Religion of Statistics as Practiced in Medical Journals," *The American Statistician*, 39, 220–223.

Schervish, M. (1996), "*P*-Values: What They are and What They are Not," *The American Statistician*, 50, 203–206.

Schmidt, F.L. (1996), "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers," *Psychological Methods*, 1, 115–129.

Schmidt, F.L., and Hunter, J.E. (2014), *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings* (3rd ed.), Thousand Oaks, CA: Sage.

Sellke, T.M., Bayarri, M.J., and Berger, J.O. (2001), "Calibration of *p* Values for Testing Precise Null Hypotheses," *The American Statistician*, 55, 62–71.

Senn, S.J. (2001), "Two Cheers for *P*?Values," *Journal of Epidemiology and Biostatistics*, 6, 193–204.

——— (2002a), Letter to the Editor re: Goodman 1992, *Statistics in Medicine*, 21, 2437–2444.

——— (2002b), "Power is Indeed Irrelevant in Interpreting Completed Studies," *BMJ*, 325, 1304.

Smith, A.H., and Bates, M. (1992), "Confidence Limit Analyses Should Replace Power Calculations in the Interpretation of Epidemiologic Studies," *Epidemiology*, 3, 449–452.

Sterne, J.A.C. (2009), *Meta-Analysis: An Updated Collection from the Stata Journal*, College Station, TX: Stata Press.

Sterne, J.A.C., and Davey Smith, G. (2001), "Sifting the Evidence—What's Wrong with Significance Tests?" *British Medical Journal*, 322, 226–231.

Stigler, S.M. (1986), *The History of Statistics*, Cambridge, MA: Belknap Press.

Thompson, B. (2004), "The 'significance' Crisis in Psychology and Education," *The Journal of Socio-Economics*, 33, 607–613.

Thompson, W.D. (1987), "Statistical Criteria in the Interpretation of Epidemiologic Data," *American Journal of Public Health*, 77, 191–194.

Trafimow, D., and Marks, M. (2015), Editorial, *Basic and Applied Social Psychology*, 37, 1–2.

Wagenmakers, E.-J. (2007), "A Practical Solution to the Pervasive Problem of *p* Values," *Psychonomic Bulletin Review*, 14, 779–804.

Walker, A.M. (1986), "Reporting the Results of Epidemiologic Studies," *American Journal of Public Health*, 76, 556–558.

Ware, J.H., Mosteller, F., and Ingelfinger, J.A. (2009), "*p*-Values," in *Medical Uses of Statistics* (3rd ed.), Bailar, J.C. and Hoaglin, D.C. (eds.), Hoboken, NJ: Wiley, pp. 175–194.

Weinberg, C.R. (2001), "Its Time to Rehabilitate the *P*-Value," *Epidemiology*, 12, 288–290.

Whitehead, A. (2002), *Meta-Analysis of Controlled Clinical Trials*, New York: Wiley.

Wood, J., Freemantle, N., King, M., and Nazareth, I. (2014), "Trap of Trends to Statistical Significance: Likelihood of Near Significant *P* Value Becoming More Significant with Extra Data," *BMJ*, 348, g2215, doi:10.1136/bmj.g2215.

Xu, L., Freeman, G., Cowling, B.J., and Schooling, C.M. (2013), "Testosterone Therapy and Cardiovascular Events Among Men: A Systematic Review and Meta-analysis of Placebo-Controlled Randomized Trials," *BMC Med.*, 11, 108.

Yates, F. (1951), "The Influence of *Statistical Methods for Research Workers* on the Development of the Science of Statistics," *Journal of the American Statistical Association*, 46, 19–34.

You ,B., Gan, H.K., Pond, G., and Chen, E.X. (2012), "Consistency in the Analysis and Reporting of Primary End Points in Oncology Randomized Controlled Trials from Registration to Publication: A Systematic Review," *Journal of Clinical Oncology*, 30, 210–216.

Ziliak, S.T., and McCloskey, D.N. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*, Ann Arbor: U Michigan Press.

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).

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.

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-reproduciblereseachrecommendations.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.