AMERICAN
Scientist

FEATURE ARTICLE

The Statistical Crisis in Science

Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

Andrew Gelman, Eric Loken

There is a growing realization that reported "statistically significant" claims in scientific publications are routinely mistaken. Researchers typically express the confidence in their data in terms of *p*-value: the probability that a perceived result is actually the result of random variation. The value of *p* (for "probability") is a way of measuring the extent to which a data set provides evidence against a so-called null hypothesis. By convention, a *p*-value below 0.05 is considered a meaningful refutation of the null hypothesis; however, such conclusions are less solid than they appear.

The idea is that when *p* is less than some prespecified value such as 0.05, the null hypothesis is rejected by the data, allowing researchers to claim strong evidence in favor of the alternative. The concept of *p*-values was originally developed by statistician Ronald Fisher in the 1920s in the context of his research on crop variance in Hertfordshire, England. Fisher offered the idea of *p*-values as a means of protecting researchers from declaring truth based on patterns in noise. In an ironic twist, *p*-values are now often used to lend credence to noisy claims based on small samples.

enlarge image

In general, *p*-values are based on what would have happened under other possible data sets. As a hypothetical example, suppose a researcher is interested in how Democrats and Republicans perform differently in a short mathematics test when it is expressed in two different contexts, involving either healthcare or the military. The question may be framed nonspecifically as an investigation of possible associations between party affiliation and mathematical reasoning across contexts. The null hypothesis is that the political context is irrelevant to the task, and the alternative hypothesis is that context matters and the difference in performance between the two parties would be different in the military and healthcare contexts.

At this point a huge number of possible comparisons could be performed, all consistent with the researcher's theory. For example, the null hypothesis could be rejected (with statistical significance) among men and not among women—explicable under the theory that men are more ideological than women. The pattern could be found among women but not among men—explicable under the theory that women are more sensitive to context than men. Or the pattern could be statistically significant for neither group, but the difference could be significant (still fitting the theory, as described above). Or the effect might only appear among men who are being questioned by female interviewers.

We might see a difference between the sexes in the healthcare context but not the military context; this would make sense given that health care is currently a highly politically salient issue and the military is less so. And how are independents and nonpartisans handled? They could be excluded entirely, depending on how many were in the sample. And so on: A single overarching research hypothesis—in this case, the idea that issue context interacts with political partisanship to affect mathematical problem-solving skills—corresponds to many possible choices of a decision variable.

This *multiple comparisons* issue is well known in statistics and has been called "*p* -hacking" in an influential 2011 paper by the psychology researchers Joseph Simmons, Leif Nelson, and Uri Simonsohn. Our main point in the present article is that it is possible to have multiple potential comparisons (that is, a data analysis whose details are highly contingent on data, invalidating published *p*-values) without the researcher performing any conscious procedure of fishing through the data or explicitly examining multiple comparisons.

How to Test a Hypothesis

In general, we can think of four classes of procedures for hypothesis testing: (1) a simple classical test based on a unique test statistic, T, which when applied to the observed data yields $T(y)$, where y represents the data; (2) a classical test prechosen from a set of possible tests, yielding $T(y;\varphi)$, with preregistered φ (for example, φ might correspond to choices of control variables in a regression, transformations, the decision of which main effect or interaction to focus on); (3) researcher degrees of freedom without fishing, which consists of computing a single test based on the data, but in an environment where a different test would have been performed given different data; the result of such a course is $T(y;\varphi(y))$, where the function $\varphi(\bullet)$ is observed for the data, y. It is generally considered unethical to (4) commit outright fishing, computing $T(y;\varphi_j)$ for $j=1,...J$. This would be a matter of performing J tests and then reporting the best result given the data, thus $T(y;\varphi_{best}(y))$.

It would take a highly unscrupulous researcher to perform test after test in a search for statistical significance (which could almost certainly be found at the 0.05 or even the 0.01 level, given all the options above and the many more that would be possible in a real study). The difficult challenge lies elsewhere: Given a particular data set, it can seem entirely appropriate to look at the data and construct reasonable rules for data exclusion, coding, and analysis that can lead to statistical significance. In such a case, researchers need to perform only one test, but that test is conditional on the data; hence, $T(y;\varphi(y))$, with the same effect as if they had deliberately fished for those results. As political scientists Macartan Humphreys, Raul Sanchez de la Sierra, and Peter van der Windt wrote in 2013, a researcher faced with multiple reasonable measures can think—perhaps correctly—that the one that produces a significant result is more likely to be the least noisy measure, but then decide—incorrectly—to draw inferences based on that one measure alone. In the hypothetical example presented earlier, finding a difference in the healthcare context might be taken as evidence that that is the most important context in which to explore differences.

This error carries particular risks in the context of small effect sizes, small sample sizes, large measurement errors, and high variation (which combine to give low power, hence less reliable results even when they happen to be statistically significant, as discussed by Katherine Button and her coauthors in a 2013 paper in *Nature Reviews: Neuroscience*). Multiplicity is less consequential in settings with large real differences, large samples, small measurement errors, and low variation. To state the problem in Bayesian terms (where *p*-values are about the plausibility of the hypothesis given the data, as opposed to the other way around), any data-based claim is more plausible to the extent it is a priori more likely, and any claim is less plausible to the extent that it is estimated with more error.

There are many roads to statistical significance; if data are gathered with no preconceptions at all, statistical significance can obviously be obtained even from pure noise by the simple means of repeatedly performing comparisons, excluding data in different ways, examining different interactions, controlling for different predictors, and so forth. Realistically, though, a researcher will come into a study with strong substantive hypotheses, to the extent that, for any given data set, the appropriate analysis can seem evidently clear. But even if the chosen data analysis is a deterministic function of the observed data, this does not eliminate the problem posed by multiple comparisons.

Arm Strength and Economic Status

In 2013, a research group led by Michael Petersen of Aarhus University published a study that claimed to find an association between men's upper-body strength, interacted with socioeconomic status, and their attitudes about economic redistribution. Using arm circumference as a proxy for arm strength, which was in turn serving as a proxy for fighting ability, the authors argue that stronger men of high socioeconomic status (SES) will oppose wealth redistribution, and that stronger men with low SES will support redistribution.

These researchers had enough degrees of freedom for them to be able to find any number of apparent needles in the haystack of their data—and, again, it would be easy enough to come across the statistically significant comparisons without "fishing" by simply looking at the data and noticing large differences that are consistent with their substantive theory.

Most notably, the authors report a statistically significant interaction with no statistically significant main effect—that is, they did not find that men with bigger arm circumference had more conservative positions on economic redistribution. What they found was that the correlation of arm circumference with opposition to redistribution of wealth was higher among men of high socioeconomic status. Had they seen the main effect (in either direction), they could have come up with a theoretically justified explanation for that, too. And if there had been no main effect and no interaction, they could have looked for other interactions. Perhaps, for example, the correlations could have differed when comparing students with or without older siblings?

As we wrote in a 2013 critique for *Slate,* nothing in this report suggests that fishing or *p* -hacking—which could imply an active pursuit of statistical significance—was involved at all. Of course, it is reasonable for scientists to refine their hypotheses in light of the data. When the desired pattern does not show up as a main effect, it makes sense to look at interactions. (For example, our earlier mention of older siblings was no joke: Family relations are often taken to be crucial in evolutionary psychology–based explanations.)

There also appear to be some degrees of freedom involved in the measurement, for example in the procedures for comparing questionnaires across countries. In conducting follow-up validations, the researchers found that some of the Danish questions worked differently when answered by Americans, and further explain: "When these two unreliable items are removed … the interaction effect becomes significant. …The scale measuring support for redistribution in the Argentina sample has a low α-level [an index of measurement precision] and, hence, is affected by a high level of random noise. Hence, the consistency of the results across the samples is achieved in spite of this noise." They conclude that "a subscale with an acceptable $\alpha = 0.65$ can be formed" from two of the items. These may be appropriate data analytic decisions, but they are clearly data dependent.

In 2013, psychologists Brian Nosek, Jeffrey Spies, and Matt Motyl posted an appealing example of prepublication replication in one of their own studies, in which they performed an experiment on perceptual judgment and political attitudes, motivated and supported by substantive theory. In their so-called 50 shades of gray study, Nosek and his coauthors found a large and statistically significant relationship between political extremism and the perception of images as black or white rather than intermediate shades. But rather than stopping there, declaring victory, and publishing these results, they gathered a large new sample and performed a replication with predetermined protocols and data analysis. According to an analysis based on their initial estimate, the replication had a 99 percent chance of reaching statistical significance with $p < 0.05$. In fact, though, the attempted replication was unsuccessful, with a *p*-value of 0.59.

enlarge image

Unwelcome though it may be, the important moral of the story is that the statistically significant *p*-value cannot be taken at face value—even if it is associated with a comparison that is consistent with an existing theory.

In Search of ESP

A much-discussed example of possibly spurious statistical significance is the 2011 claim of Daryl Bem, an emeritus professor of social psychology at Cornell University, to have found evidence for extrasensory perception (ESP) in college students. In his first experiment, in which 100 students participated in visualizations of images, he found a statistically significant result for erotic pictures but not for nonerotic pictures. Despite the misgivings of many critics such as the psychometrician E. J. Wagenmakers, the study was published in a prestigious journal and received much media attention. After some failed attempts at replications, the furor has mostly subsided, but this case remains of interest as an example of how investigators can use well-accepted research practices to find statistical significance anywhere.

Bem's paper presented nine different experiments and many statistically significant results—multiple degrees of freedom that allowed him to keep looking until he could find what he was searching for. But consider all the other comparisons he could have drawn: If the subjects had identified all images at a rate statistically significantly higher than chance, that certainly would have been reported as evidence of ESP. Or what if performance had been higher for the nonerotic pictures? One could easily argue that the erotic images were distracting and only the nonerotic images were a good test of the phenomenon. If participants had performed statistically significantly better in the second half of the trial than in the first half, that would be evidence of learning; if better in the first half, evidence of fatigue.

Bem, in a follow-up paper with statisticians Jessica Utts and Wesley Johnson, rebutted the criticism that his hypotheses had been exploratory. On the contrary, the three wrote, "The specificity of this hypothesis derives from several earlier 'presentiment' experiments (e.g., Radin, 1997) which had demonstrated that participants showed anomalous 'precognitive' physiological arousal a few seconds before seeing an erotic image but not before seeing a calm or nonerotic image." The authors explained they had also presented nonerotic images mixed in at random intervals with the erotic ones to leave open the question of whether the participants could anticipate the future left/right positions of these images. They could not do so, a finding that Bem and his coauthors saw as "consistent with the results of the presentiment experiments." Summing up, they state that "there was no data exploration that required adjustment for multiple analyses in this or any other experiment."

We have no reason to disbelieve the above description of motivations, but it seems clear to us that each of the scientific hypotheses there described correspond to multiple statistical hypotheses. For example, consider the statement about "anomalous precognitive physiological arousal." Suppose that the experimental subjects had performed statistically significantly worse for the erotic pictures. This result, too, would fit right into the theory, with the rationale that the anomalous arousal could be interfering with otherwise effective precognitive processes.

Bem insists his hypothesis "was not formulated from a post hoc exploration of the data," but a data-dependent analysis would not necessarily look "post hoc." For example, if men had performed better with erotic images and women with romantic but nonerotic images, there is no reason such a pattern would look like fishing or *p* -hacking. Rather, it would be seen as a natural implication of the research hypothesis, because there is a considerable amount of literature suggesting sex differences in response to visual erotic stimuli. The problem resides in the one-to-many mapping from scientific to statistical hypotheses.

Menstrual Cycles and Voting

For a *p*-value to be interpreted as evidence, it requires a strong claim that the same analysis would have been performed had the data been different. In 2013, psychologists Kristina Durante, Ashley Rae, and Vladas Griskevicius published a paper based on survey data claiming that "Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led married women to become more conservative, more religious, and more likely to vote for Mitt Romney.…Overall, the ovulatory cycle not only influences women's politics, but appears to do so differently for single versus married women." The claimed effects were huge, indeed implausibly large given our understanding of the stability of political partisanship: for example, they report that, among women in relationships, 40 percent in the ovulation period supported Romney, compared to 23 percent in the nonfertile part of their cycle.

But the reported comparison was statistically significant: Does that mean we are duty-bound to believe it, or at least to consider the data as strong evidence in favor of their hypothesis? No, and the reason is, again, the garden of forking paths: Even if Durante and her colleagues only performed one analysis on the particular data they saw, had they seen other data, they could have performed other analyses that would be equally consistent with their substantive theory.

The interaction reported in the paper (a different pattern for married and single women) coheres with the authors' general theoretical perspective ("ovulation should lead women to prioritize securing genetic benefits from a mate possessing indicators of genetic fitness"). But various other main effects and interactions would also fit the theory. Indeed, as the authors note, their hypothesis "is consistent with the idea that women should support the more liberal candidate." Or suppose the data had followed the opposite pattern, with time of ovulation (as estimated by the researchers) being correlated with conservative attitudes among single women and with liberal attitudes among married women. This would fit a story in which ovulation leads women's preferences away from party identification and toward more fundamental biological imperatives. Other natural interactions to consider would be age or socioeconomic status (as in the arm-circumference paper considered earlier).

On a first reading, these objections may seem petty. After all, these researchers found a large effect that was consistent with their theory, so why quibble if the significance level was somewhat overstated because of multiple comparisons problems? We believe it is important to call attention to these flaws, however, for two reasons. First, the claimed effect size, in the range of a 20 percentage point difference in vote intention at different phases of the menstrual cycle, is substantively implausible, given all the evidence from polling that very few people change their vote intentions during presidential general election campaigns (a well-known finding that Gelman and colleagues recently confirmed with a panel survey from the 2012 presidential election campaign). Second, the statistical significance of the published comparisons is a central part of the authors' argument—certainly the paper would not have been published in a top journal without $p < 0.05$ results—and the high multiplicity of all the potential interactions is relevant to this point.

Would the same data-analysis decisions have been made with a different data set?

In addition to the choice of main effects or interactions, Durante and her collaborators had several political questions to work with (attitudes as well as voting intentions), along with other demographic variables (age, ethnicity, and parenthood status) and flexibility in characterizing relationship status (at one point, "single" versus "married," but later, "single" versus "in a committed relationship").

Data Processing and Data Analysis

We have considered several prominent research papers in which statistical significance was attained via a sort of invisible multiplicity: data-dependent analysis choices that did not appear to be degrees of freedom because the researchers analyze only one data set at a time. Another study, also published in a top psychology journal, exhibits several different forms of multiplicity of choices in data analysis.

In 2013, psychologists Alec Beall and Jessica Tracy reported in *Psychological Science* that women who were at peak fertility were three times more likely to wear red or pink shirts than women at other points in their menstrual cycles. The researchers' theory, they wrote, was "based on the idea that red and shades of red (such as the pinkish swellings seen in ovulating chimpanzees, or the pinkish skin tone observed in attractive and healthy human faces) are associated with sexual interest and attractiveness." In a critique published later that year in *Slate*, one of us (Gelman) noted that many different comparisons could have been reported in the data, so there was nothing special about a particular comparison being statistically significant.

Tracy and Beall responded on the website of their Emotion and Self Lab at the University of British Columbia that they had conducted their studies "with the sole purpose of testing one specific hypothesis: that conception risk would increase women's tendency to dress in red or pink"—a hypothesis that they saw as emerging clearly from a large body of work, which they cited. "We set out to test a specific theory," they write.

Nevertheless, it seems clear to us that their analysis was contingent on the data: Within the context of their specific theory are many possible choices of data selection and analysis. Most important, their protocol and analysis were not preregistered. Even though Beall and Tracy did an analysis that was consistent with their general research hypothesis—and we take them at their word that they were not conducting a "fishing expedition"—many degrees of freedom remain in their specific decisions: how strictly to set the criteria regarding the age of the women included, the hues considered as "red or shades of red," the exact window of days to be considered high risk for conception, choices of potential interactions to examine, whether to combine or contrast results from different groups, and so on.

Again, all the above could well have occurred without it looking like *p* -hacking or fishing. Rather, the researchers start with a somewhat formed idea in their mind of what comparison to perform, and they refine that idea in light of the data. This example is particularly stark because Beall and Tracy on one hand, and Durante and her coauthors on the other, published two studies inspired by similar stories, using similar research methods, in the same journal in the same year. But in the details they made different analytic choices, each time finding statistical significance with the comparisons they chose to focus on. Both studies compared women in ovulation and elsewhere in their self-reported menstrual cycles, but they used different rules for excluding data and different days for their comparisons. Both studies examined women of childbearing age, but one study reported a main effect whereas the other reported a difference between single and married women. In neither case were the data inclusion rules and data analysis choices preregistered.

In this garden of forking paths, whatever route you take seems predetermined, but that's because the choices are done implicitly. The researchers are not trying multiple tests to see which has the best *p*-value; rather, they are using their scientific common sense to formulate their hypotheses in a reasonable way, given the data they have. The mistake is in thinking that, if the particular path that was chosen yields statistical significance, this is strong evidence in

favor of the hypothesis.

Criticism is Easy, Research is Hard

Flaws can be found in any research design if you look hard enough. Our own applied work is full of analyses that are contingent on data, yet we and our colleagues have been happy to report uncertainty intervals (and thus, implicitly, claims of statistical significance) without concern for selection bias or multiple comparisons. So we would like to put a positive spin on the message of this paper, to avoid playing the role of statistician as scold.

In our experience, it is good scientific practice to refine one's research hypotheses in light of the data. Working scientists are also keenly aware of the risks of data dredging, and they use confidence intervals and *p*-values as a tool to avoid getting fooled by noise. Unfortunately, a by-product of all this struggle and care is that when a statistically significant pattern does show up, it is natural to get excited and believe it. The very fact that scientists generally don't cheat, generally don't go fishing for statistical significance, makes them vulnerable to drawing strong conclusions when they encounter a pattern that is robust enough to cross the $p < 0.05$ threshold.

We are hardly the first to express concern over the use of *p*-values to justify scientific claims, or to point out that multiple comparisons invalidate *p*-values. Our contribution is simply to note that because the justification for *p*-values lies in what would have happened across multiple data sets, it is relevant to consider whether any choices in analysis and interpretation are data dependent and would have been different given other possible data. If so, even in settings where a single analysis has been carried out on the given data, the issue of multiple comparisons emerges because different choices about combining variables, inclusion and exclusion of cases, transformations of variables, tests for interactions in the absence of main effects, and many other steps in the analysis could well have occurred with different data. It's also possible that different interpretations regarding confirmation of theories would have been invoked to explain different observed patterns of results.

The issue of multiple comparisons arises even with just one analysis of the data.

At this point it might be natural to object that any research study involves data-dependent decisions, and so is open to the critique outlined here. In some sense, yes. But we have discussed examples where we find a strong reliance on the *p*-value to support a strong inference. In the case of the ESP experiments, a phenomenon with no real theoretical basis was investigated with a sequence of studies designed to reveal small effects. The studies of women's voting behavior, men's attitudes about the distribution of wealth, and women's tendency to wear red when ovulating, are all broadly consistent with evolutionary theories, but produced implausibly large effects in relatively small studies.

What, then, can be done?

In political science, Humphreys and his coauthors recommend preregistration: defining the entire data-collection and analysis protocol ahead of time. For most of our own research projects this strategy hardly seems possible: In our many applied research projects, we have learned so much by looking at the data. Our most important hypotheses could never have been formulated ahead of time. For example, one of Gelman's most successful recent projects was a comparison of the attitudes of rich and poor voters in rich and poor states; the patterns found by Gelman and his collaborators became apparent only after many different looks at the data (although they were confirmed by analyses of other elections). In any case, as applied social science researchers we are often analyzing public data on education trends, elections, the economy, and public opinion that have already been studied by others many times before, and it would be close to meaningless to consider preregistration for data with which we are already so familiar.

In fields such as psychology where it is typically not so difficult to get more data, preregistration might make sense. At the same time, we do not want demands of statistical purity to strait-jacket our science, whether in psychology, nutrition, or education. The most valuable statistical analyses often arise only after an iterative process involving the data. Preregistration may be practical in some fields and for some types of problems, but it cannot realistically be a general solution.

One message we wish to emphasize is that researchers can and should be more aware of the choices involved in their data analysis, partly to recognize the problems with published *p*-values but, ultimately, with the goal of recognizing the actual open-ended aspect of their projects and then analyzing their data with this generality in mind. One can follow up an open-ended analysis with prepublication replication, which is related to the idea of external validation, popular in statistics and computer science. The idea is to perform two experiments, the first being exploratory but still theory-based, and the second being purely confirmatory with its own preregistered protocol.

In (largely) observational fields such as political science, economics, and sociology, replication is difficult or infeasible. We cannot easily gather data on additional wars, or additional financial crises, or additional countries. In such settings our only recommendation can be to more fully analyze existing data. A starting point would be to analyze all relevant comparisons, not just focusing on whatever happens to be statistically significant. We have elsewhere argued that multilevel modeling can resolve multiple-comparisons issues, but the practical difficulties of such an approach are not trivial.

The Way Forward

We must realize that, absent preregistration or opportunities for authentic replication, our choices for data analysis will be data dependent, even when they are motivated directly from theoretical concerns. When preregistered replication is difficult or impossible (as in much research in social science and public health), we believe the best strategy is to move toward an analysis of all the data rather than a focus on a single comparison or small set of comparisons. There is no statistical quality board that could enforce such larger analyses—nor would we believe such coercion to be appropriate—but as more and more scientists follow the lead of Brian Nosek, who openly expressed concerns about the malign effects of *p*-values on his own research, we hope there will be an increasing motivation toward more comprehensive data analyses that will be less subject to these concerns. If necessary, one must step back to a sharper distinction between exploratory and confirmatory data analysis, recognizing the benefits and limitations of each.

In fields where new data can readily be gathered, perhaps the two-part structure of Nosek and his colleagues—attempting to replicate his results before publishing—will set a standard for future research. Instead of the current norm in which several different studies are performed, each with statistical significance but each with analyses that are contingent on data, perhaps researchers can perform half as many original experiments in each paper and just pair each new experiment with a preregistered replication. We encourage awareness among scientists that *p*-values should not necessarily be taken at face value. However, this does not mean that scientists are without options for valid statistical inference.

Our positive message is related to our strong feeling that scientists are interested in getting closer to the truth. In the words of the great statistical educator Frederick Mosteller, it is easy to lie with statistics, but easier without them.

+ enlarge image

Bibliography

- Beall, A. T., and J. L. Tracy. 2013. Women are more likely to wear red or pink at peak fertility. *Psychological Science* 24(9):1837–1841.
- Bem, D. J. 2011. Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology* 100:407–425.
- Bem, D. J., J. Utts, and W. O. Johnson. 2011. Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology* 101:716–719.
- Box, G. E. P. 1997. Scientific method: The generation of knowledge and quality. *Quality Progress* January: 47–50.
- Button, K. S., et al. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews: Neuroscience* 14:365–376.
- de Groot, A. D. 1956. The meaning of the concept of significance in studies of various types. *Nederlands Tijdschrift voor de Psychologie en Haar Grensgebieden* 11:398–409. Translated in 2013 by E. J. Wagenmakers et al. https://dl.dropboxusercontent.com/u/1018886/Temp/DeGroot v3.pdf
- Durante, K., A. Rae, and V. Griskevicius. 2013 The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science* 24:1007–1016.
- Gelman, A. 2013. Is it possible to be an ethicist without being mean to people? *Chance* 26.4:51–53 .
- Gelman, A. 2014. The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management* .
- Gelman, A., J. Hill, and M. Yajima. 2012. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness* 5:189–211.
- Humphreys, M., R. Sanchez, and P. Windt. 2013. Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis* 21:1–20.
- Nosek, B. A., J. R. Spies, and M. Motyl. 2013. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science* 7:615–631.
- Petersen, M. B., D. Sznycer, A. Sell, L. Cosmides, and J. Tooby. 2013. The ancestral logic of politics: Upper-body strength regulates men's assertion of self-interest over economic redistribution. *Psychological Science.* http://dx.doi.org/10.2139/ssrn.1798773
- Simmons, J., L. Nelson, and U. Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science* 22:1359–1366.
- Tracy, J. L., and A. T. Beall. 2013. Too good does not always mean not true. University of British Columbia Emotion and Self Lab, July 30. http://ubc-emotionlab.ca/2013/07/too-good-does-not-always-mean-not-true/
- Wagenmakers, E. J., R. Wetzels, D. Borsboom, and H. L. J. van der Maas. 2011. Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem. *Journal of Personality and Social Psychology* 100:426–432.