Home
News and ideas
Managing my journal
Peer review
Raising the profile of my journal
Citations, impact and usage
Ethics and rights
Open Access (OA)
Taylor & Francis Online

Latest from Twitter

"Tweets announcing the publication of papers made a serious difference" Read about **@Altmetric #socialmedia** & impact: **https://t.co/P2auUb6dsj**

Everyone is reading our open access fact sheets this week! Everything your authors need to know in just 5 minutes: **https://t.co/iIDIx39Wiq**

Popular article tags

November 16, 2015 | *Amanda Ashworth, Publisher*

Veto on the use of null hypothesis testing and p intervals: right or wrong?

It's a brave editor who takes a decision to change accepted practice for submissions and peer review, particularly when he knows that his reasoning is controversial, that there are strong opposing views, and the reaction from the scholarly community is likely to be highly polarised – and very vocal. But that didn't stop Dr. David Trafimow, editor of *Basic and Applied Social Psychology*, from announcing in an editorial in the first issue of 2015 that the journal will cease accepting papers that relied on certain statistical methods – especially the null hypothesis significance testing procedure (NHSTP) – with immediate effect.

Since publication of the editorial, both the journal and its Editor-in-Chief have received a huge amount of attention. The editorial has been viewed over 100,000 times to date and has attracted a number of tweets, posts, blog entries, and coverage in publications such as *Scientific American* and the *Economist*, all discussing the merit and implications of the journal's decision.

Dr. Trafimow (of New Mexico State University) has long held that reliance on NHSTP results leads to publication of flawed research and is inappropriate for work in social psychology. Can a journal's editorial policies change the way a research field uses statistical tests? And what is it like for a journal editor to face a storm of criticism when he follows through on his convictions about what's right for science? Would he advise other editors to follow his lead? We invited Dr. Trafimow to reflect on the reaction to his editorial and what it means for the future of scholarly debate around the issue.

Dr. David Trafimow, Editor-In-Chief, Basic and Applied

Social Psychology

I was surprised at the extent of the reaction, which perhaps indicates some naïveté on my part. It would be nice to say that I have completely processed all of the reaction, but this is not so. The reaction has been so extensive that I have been able to read only a portion of it. In addition to blogs, tweets, and posts, many people have sent me emails. These include highly encouraging and supportive ones, thoughtful criticisms, questions, and others, and I feel that all of these were positive in various ways.

The 2015 editorial in *Basic and Applied Social Psychology* (BASP) is better viewed as the result of a long period of inquiry and consideration than as a sudden change in statistical thinking. There have been published objections to the null hypothesis significance testing procedure for many decades. I first wrote about this issue in 2003, in an article in the journal *Psychological Review*, and more recently in an editorial in BASP in 2014. The 2014 editorial discouraged the procedure but did not ban it. Unfortunately, the 2014 editorial did not decrease the use of null hypothesis significance tests in BASP submissions, which supported my intuition that additional measures would be needed. Hence, we adopted the stronger position you see in the 2015 editorial.

Out of the chaos that has characterized my life since the editorial came out, there is one thought that remains particularly salient. Specifically, whatever the consequences of the discussion turn out to be, the discussion itself is badly needed, and it seems that many people in a variety of sciences are seeing that. By stimulating that discussion, the editorial has made a contribution.

I believe that the debate is an extremely positive development because I hope and expect that poor arguments will not be able to withstand prolonged intense scrutiny by a variety of sharp minds. When the arguments have been examined, some may remain plausible but many will not. The result will be better practices for drawing conclusions from the findings that researchers obtain, and I consider this to be a major gain for science.

We also invited respected colleagues to respond to several key questions (below) that are central to this ongoing scholarly debate.

Dr. Susan S. Ellenberg, Professor of Biostatistics, Perelman School of Medicine, University of Pennsylvania

Question: The position taken by BASP editors covers the null hypothesis significance testing procedure but not all inferential statistics – can you comment on the scope of the position taken?

Answer: Statistical testing procedures were developed to prevent researchers from becoming overly excited about differences that could easily have been attributed to chance. If we flip a fair coin 20 times, we don't expect to get exactly 10 heads and 10 tails. We understand that the split might well be 9-11 or 8-12. It is natural to wonder, though, how much of an

imbalance should make us question whether the coin is really fair. The probabilities of extreme imbalances can be readily calculated – even a split of 2-18 will be observed (rarely) with a fair coin, and different people may have different thresholds for becoming suspicious due to observing a given imbalance. The use of p-values simply tells people how often we would observe a difference as extreme as or more extreme than we did observe, if there really were no difference. Yes, it is widely misunderstood as the probability that there is a difference, but that doesn't mean it's not a useful tool to help gauge whether a conclusion that one outcome differs from the other is reasonably reliable. A split in 20 coin flips of 2 vs 18 doesn't tell us what the probability is of the coin being fair, but few would argue that such a split doesn't provide reasonable evidence against it being a fair coin. In reports of medical research, different investigators will have their own thresholds for believing that there is a difference, versus attributing the finding to chance (or to possible confounding factors, problems with the quality of study conduct, selective reporting, etc.), but providing some quantification of the risk of erroneous conclusions (or, as with Bayesian methods, the probability of having made a correct inference) is an important starting point.

So, what is the problem? The problem is that for many investigators, a p-value below a certain value has come to be seen as the "holy grail," with research not achieving the desired level of significance being regarded as completely uninformative. I believe some journal editors take this perspective also, being much more likely to accept papers with results reaching conventional levels of statistical significance, leading to a well-established bias that frustrates those who perform meta-analyses based on published data.

The solution that Dr. Trafimow has proposed, to do away with reporting of p-values and focus only on "descriptive statistics," is wrong-headed. He wishes to remove a tool that helps both investigators and journal readers to assess the role of chance in the reported findings. A much better solution for a journal editor is to develop a culture in which reviewers and associate editors ensure that reported significance levels are used and interpreted appropriately. This would require authors to indicate what hypotheses were pre-specified, what outcomes were actually tested, whether any observations were discarded and why – all the kinds of things that investigators may do that make the p-values that they report uninterpretable. Many journals now require that authors provide this information.

Dr. David L. Streiner, Professor, Department of Psychiatry, University of Toronto

Question: To what extent is the null hypothesis significance procedure flawed and how much of the problem is simply that researchers misuse the procedure?

Answer: BASP has taken the position that henceforth all vestiges of the null hypothesis testing procedure (NHSTP) will be banned from its pages. This includes values of p, t, F, confidence intervals, and statements about

significance. In their place, authors are encouraged to use larger sample sizes and "strong descriptive statistics, including effect sizes." Does this make sense? I believe it does not. I expect that authors will interpret the effect sizes (ESs) by referring back to Cohen, and stating that those under 0.2 are small, 0.5 is moderate, and those 0.8 and greater are larger. These values have assumed the position of holy writ handed down on Mount Sinai, ignoring the fact that Cohen defined them hesitantly, stating that "there is a certain risk inherent in offering conventional operational definitions for those terms for use in power analysis in as diverse a field of inquiry as behavioral science" [1]. Moreover, we should bear in mind the words of Thompson, who said that "If people interpreted effect sizes with the same rigidity that $\alpha = .05$ has been used in statistical testing, we would merely be being stupid in another metric" [2]. In other words, replacing NHSTP and confidence intervals with ESs simply replaces interpreting one type of statistics as "significant" or "important" with another.

References

1. Cohen, J. (1988) *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, p. 25.

2. Thompson, B. (2001) Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education*. 70 (1). 80–93.

Mr. Robert Grant, Senior Lecturer in Health & Social Care Statistics, Kingston University & St. George's, University of London

Question: How do you see this debate evolving, and how would you like to see it evolve?

Answer: Although I feel that an outright ban of hypothesis tests constitutes throwing the baby out with the bathwater, it does have the very positive effect of stimulating discussion on this subject of the goals of, and best practice in, scientific data analysis. These issues are considered not nearly as much as they ought to be, and yet ironically the problems that have provoked the ban are well known and have been written about eloquently, if one knows where to look. I would like briefly to consider three statistical aspects: hypothesis testing, confidence intervals, and Bayesian methods, then move on to reflect on the crucial role of education and training for researchers. I am a medical statistician, and my examples are necessarily tilted in that direction, but the methodological considerations are universal.

Strictly speaking, there is no debate. We know that hypothesis testing in the Neyman-Pearson framework is mathematically sound and logically consistent, and anyone who doubts it can plough through the theorems, proofs, and lemmas in any theoretical statistics textbook. However, a lot of what we study now does not lend itself to Neyman & Pearson's view of a simple and quantifiable world, built on the experience of early twentieth-century agricultural and industrial experiments: make mutually exclusive

hypotheses under demonstrable probability assumptions, state a loss function for errors, and derive a test that minimizes the loss function. Almost nobody does that, but it is very common to see the tests that arose from this framework done mechanistically without an understanding of why we might do them. As Andrew Gelman has written, the trouble with p-values is the way they are used [1]. If we regard them as providing just another summary of the data, this time of how convincing a pattern is under a certain probability model for sampling, then we acknowledge that we still have some interpretation to do after finding p-values. They become much safer, but also harder work and open to some more researchers' (and readers') biases.

Confidence intervals are not immune to the problem of overblown conclusions based on which side of an arbitrary line the null value falls. A fascinating review recently compared three high-impact, peer-reviewed medical meta-analyses, which might be hoped to form the gold standard of research conduct, looking at almost identical studies, getting almost identical results, and drawing wildly different conclusions: conclusions that influence how doctors would treat serious illness [2]. We have not advanced much from the days of medicine as art, not science. So, like hypothesis tests, we need to be open about our interpretation of findings in the context of the study and our mental causal models of how the data came about.

And this brings us to Bayes. It is widely believed that Bayes is synonymous with subjectivity and probability as opinion. While this interpretation of probability is, like Neyman-Pearson, mathematically sound, many who would count themselves as Bayesians remain unconvinced of the utility of subjective probability in communicating research findings to humans (machine learning is another matter). I count myself in this camp. The growth in popularity of Bayesian methods in recent years owes more to the flexibility with which it can solve difficult problems, and availability of user-friendly software, rather than any decline of positivism. In fact, the only essential feature distinguishing Bayesian from frequentist methods is that probabilities can be applied to anything we do not know, rather than things which can be identically and eternally replicated. This is the distinction written on the first page of Bayes's posthumous paper for the Royal Society, and the requirements of frequentism quickly lead to contradictions, leaving Bayes as the only good contender. With this in mind, surely it is better to acknowledge our uncertainties and biases and incorporate them in the calculation where possible, rather than brush them under the carpet. For that reason, I welcome the promotion of Bayesian methods in the editorial.

I believe this debate should broaden to consider how we can improve the status quo, particularly through teaching introductory statistics. There is a further difficult problem of performance indicators in academia, which incentivize bold and certain conclusions, but that is under attack already, so I will focus on teaching. At present, students of a first course in data analysis learn the bad habit of cookbook statistics from the beginning, and this leads to the mechanistic use of statistics to draw firm conclusions. This is my theory about it in psychology (I hope it does not offend): through decades of (unfairly) not being taken seriously as a science, psychologists acquired the habit of foregrounding the trappings of serious data analysis to

demonstrate their credentials. Indeed, psychologists were responsible for many advances in multivariate analysis which are widely used. But with time, these techniques have hardened into a ritualized procedure that discourages budding researchers from understanding what and why they are actually calculating, while offering the false certainty of a recognized set of steps that lead to answers.

We should spend less time teaching the recipes to do particular calculations and tests, and focus instead on the ability to conceive clearly and precisely of the problem and question at hand. With relevant computing skills and a little practice at finding tools to address the problem at hand, researchers will then be equipped to tackle real-life problems while recognizing their own limits. Thankfully, there is a growing momentum behind these principles as set out in the American Statistical Association's Guidelines for Assessment and Instruction in Statistics Education [3].

References

1. Gelman, A. Columbia University. (2013) *The problem with p-values is how they're used. [Online] Available from:* http://www.stat.columbia.edu/~gelman/research/unpublished/murtaugh2.p

2. McCormack, J., Vandermeer, B. and Allan G.M. (2013) How confidence intervals become confusion intervals. *BMC Medical Research* [Online] Biomedcentral. Available from: http://www.biomedcentral.com/1471-2288/13/134.

3. Garfield J, et al. GAISE College Report. American Statisrical Association. (2005) GAISE College Report. [Online] Available from: http://www.amstat.org/education/gaise/GaiseCollege_Full.pdf.

Dr. James W. Grice, Department of Psychology, Oklahoma State University

Question: What do you feel are the implications of the journal's policy for social psychologists and the field more generally?

Answer: The implications for social psychologists and psychology in general can be viewed negatively or positively depending upon which meaning of the word "implication" is being considered. In one sense of the word, the ban indicates that social psychologists, like so many other psychologists, sociologists, medical researchers, and biologists, have been implicated in a research methodology that has been known to be highly problematic since at least the 1960s. In his book *On Method*, published in 1967, David Bakan wrote a thoughtful critique of what we now generically refer to as Null Hypothesis Significance Testing, pointing out that it was entirely insufficient as a tool for judging the scientific importance of one's findings. Many notable psychologists such as Paul Meehl, David Lykken, and Jacob Cohen followed in support of Bakan's central thesis in an effort to end psychologists' obsession with NHSTP. Alongside these published critiques were surveys that revealed majorities of psychologists possessed erroneous

views of what a statistically significant result even means. In other words, they could not accurately define the meaning of "$p < .05$." The American Psychological Association was finally prompted to convene a task force in the 1990s to offer recommendations for loosening NHSTP's stranglehold on research practice (see Grice, 2011, for a summary of this history). Despite all of these efforts, the vast majority of psychologists simply went about their business publishing papers in which they used NHSTP as the primary tool for evaluating their findings. Consequently, I view Dr. Trafimow's banning of NHSTP as something of an indictment of us all because we are all implicated in perpetuating behavior that is clearly unbecoming of scientists. With all of the problems with NHSTP, why has it taken the brave actions of an editor to change our behaviors? How could we fail to do the right thing by not getting beyond the p -value on our own?

In another sense of the word, the implication of the NHSTP ban is that social psychologists must now search for meaningful and fruitful ways forward. Possible avenues of progress include the so-called "new statistics" (Cumming, 2012) and Bayesian statistics. I must admit, however, that I am not overly enthusiastic about these methods. The new statistics can be viewed as essentially an extension of the Fisher-Neyman-Pearson hybrid that is today's NHSTP, and Gigerenzer and Marewski (2015) have argued that the behavior of researchers using Bayesian statistics is not fundamentally different from the behavior of researchers using NHSTP. From my point of view, a more promising route forward is to move from the culture of parameter estimation and model fitting to the culture of pattern analysis and algorithmic modeling, as described in Breiman's (2001) article, *Statistical modeling: The two cultures*. Social psychologists often desire to open the "black box" of the psyche and understand the hidden causes behind behavior. Recent reports of extreme publication bias (see Yong, 2012) and the lack of replicable results in psychology (Open Science Collaboration, 2015), however, show clearly that traditional statistics and NHSTP are not yielding scientific knowledge, which is knowledge of causes and effects. As I have written elsewhere (Grice, 2014), greater nuance in our understanding of causality will be necessary to move forward, as will a serious commitment to an honest and open discussion of the measurement crisis in psychology. The implication, or conclusion, to be drawn from Dr. Trafimow's ban on NHSTP is therefore to fundamentally re-evaluate our ways of thinking about and studying psychological phenomena so that we may open up new avenues of thought. This move will also force us to develop innovative research methods and data analytic tools that are optimally suited for studying the human person.

References

1. Bakan, D. (1967) *On method: Toward a reconstruction of psychological investigation*. San Francisco: Jossey-Bass.

2. Breiman, L. (2001) Statistical modeling: The two cultures. *Statistical Science. 16* (3). pp. 199–231.

3. Cumming, G. (2012) *Understanding The New Statistics: Effect Sizes,*

Confidence Intervals, and Meta-Analysis. New York: Routledge.

4. Gigerenzer, G. and Marewski, J. N. (2015) Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, 41(2). pp. 421–440.

5. Grice, J. W. (2011) *Observation Oriented Modeling: Analysis of Cause in the Behavioral Sciences*. New York, NY: Academic Press.

6. Grice, J. W. (2014) Observation oriented modeling: Preparing students for research in the 21st *Innovative Teachin.3 (3).*

7. Open Science Collaboration. (2015) Estimating the reproducibility of psychological science. *Science, 349*, Issue 6251.

8. Yong, Ed. (2012) Replication studies: Bad copy. *Vol. 485*, no. 7398. pp. 298–300.

Because there is a plethora of views surrounding this debate, we invited Dr. Trafimow to reflect on these scholars' reactions to these questions as well.

Dr. David Trafimow, Editor-In-Chief, Basic and Applied Social Psychology

Before responding to the comments, I would like to thank the people who made them and also Amanda Ashworth and Adam Burbage at Taylor & Francis for initiating the project.

I'll commence by respectfully disagreeing with Dr. Ellenberg's statement that p-values are "a tool that helps both investigators and journal readers to assess the role of chance in the reported findings." I beg to differ. A p-value is a conditional probability of obtaining the finding (or one more extreme) given that the null hypothesis is true. Because one does not know whether the null hypothesis is true, there is no way to calculate the probability that the finding is due to chance. Nor do p-values permit a logically valid inference about the probability that the null hypothesis is true or false. Dr. Ellenberg also states that p-values provide a "useful tool to help gauge whether a conclusion that one outcome differs from the other is reasonably reliable." Again, this simply is not so; p-values are only slightly related to probabilities of replication and the recent Open Science Collaboration, to which Dr. Grice referred, showed a dismal rate of replicating "statistically significant" studies based on p-values.

Dr. Streiner worries about whether the p-value ban will simply result in arbitrary cut-offs based on effect sizes rather than p-values. I sympathize with this worry but consider the following. First, by arguing that the p-value ban will result in the replacement of one undesirable practice with another undesirable practice, Dr. Streiner tacitly admits that the present practice is undesirable. Second, there is no reason why eliminating an undesirable practice has to result in its replacement with another undesirable practice.

Sometimes the elimination of undesirable practices can result in better ones replacing them.

I largely agree with Dr. Grant, especially his statement that confidence intervals fail to solve the problem and his call for better education. But I respectfully disagree with the statement that the p-value ban "constitutes throwing the baby out with the bathwater." The only statement Dr. Grant makes about the value of p-values is "as providing just another summary of the data, this time of how convincing a pattern is under a certain probability model for sampling." Yet, Dr. Grant admits that p-values are insufficient for drawing conclusions. Certainly, p-values do not provide the probability that the findings are due to chance or the probability of a hypothesis or the probability of replication or the probability of anything else that actually is useful to scientists. The p-value ban did not constitute throwing out the baby with the bathwater because there is no baby. Dr. Grice correctly emphasizes that the invalidity of the null hypothesis significance testing procedure has been pointed out many times in the history of psychology but that the procedure nevertheless continues to dominate. I also appreciate his citing the recent evidence pertaining to the replication crisis in psychology, especially the Open Science Collaboration (2015) that resulted in such alarming findings. Surely, something is rotten in the state of psychology! Most important, I underscore Dr. Grice's general conclusion about the need to "fundamentally re-evaluate our ways of thinking *about* and studying psychological phenomena so that we may open up new avenues of thought." If the p-value ban helps to accomplish this, its value will have been demonstrated.

To steal a quote from Dr. Trafimow, "Surely, something is rotten in the state of psychology!" Whether we are discussing the p-value ban or the Reproducibility Project, it seems that everyone is on board with shaping up the way social scientists, statisticians, and medical experts, to name a few, conduct and output their research. I'm sure the editorial board of *Basic and Applied Social Psychology* would agree with Tom Siegfried's latest shout-out to the journal and the ban in his article "Top 10 ways to save science from its statistical self" recently published in Science News: "Ideally there would be an act of Congress (and a United Nations resolution) condemning p-values and relegating them to the same category as chemical weapons and smoking in airplanes." Now, wouldn't that be something!

Published: November 16, 2015 | **Author:** Amanda Ashworth, Publisher | **Category:** Front page, News and ideas | **Tagged with**: Community • Editorial • Guidelines • Industry • Integrity • Peer review • Policies