Questions and Answers: Reproducibility and a Stricter Threshold for Statistical Significance

Daniel Benjamin, Justin Esarey, Blakeley B. McShane,
Jennifer L. Tackett, and E.J. Wagenmakers

December 6, 2017

“Redefine statistical significance,” a paper recently published in *Nature Human Behavior* (Benjamin et al., 2017), generated a substantial amount of discussion in methodological circles. This paper proposes to lower the α threshold for statistical significance from the conventional level of 0.05 to a new, more stringent level of 0.005 and to apply this threshold specifically to newly discovered relationships (i.e., relationships that have not yet been demonstrated in multiple studies). This proposal touched off a debate about the effect null hypothesis significance testing (NHST) has on published work in the social and behavioral sciences in which many statisticians and social scientists have participated. Some have proposed alternative reforms that they believe will be more effective at improving the replicability of published results.

To facilitate further discussion of these proposals—and perhaps to begin to develop an actionable plan for reform—the International Methods Colloquium (IMC) hosted a panel discussion on “reproducibility and a stricter threshold for statistical significance” on October 27, 2017. The one-hour discussion included six panelists and over 240 attendees, with each panelist giving a brief initial statement concerning the proposal to “redefine statistical significance” and the remainder of the time being devoted to questions and answers from the audience. The event was recorded and can be viewed online for free at the International Methods Colloquium website.

Unfortunately, the IMC’s time limit of one hour prevented many audience members from asking their questions and having a chance to hear our panelists respond. Panelists and audience members alike agreed that the time limit was not adequate to fully explore all the issues raised by Benjamin et al. (2017). Consequently, questions that were not answered during the presentation were forwarded to all panelists, who were given a chance to respond.

The questions and answers, both minimally edited for clarity, are presented in this article. Questions are presented in the order they were asked; answers are presented in alphabetical order of the panelists' names. The panelists are (also in alphabetical order):

1. Daniel Benjamin, Associate Research Professor of Economics at the University of Southern California and a primary co-author of the paper in *Nature Human Behavior* (Benjamin et al., 2017) as well as many other articles on inference and hypothesis testing in the social sciences.

2. Justin Esarey, Associate Professor of Political Science at Rice University, Principal Investigator for the International Methods Colloquium and author of "Lowering the threshold of statistical significance to $p < 0.005$ to encourage enriched theories of politics" (Esarey, 2017), a response to "Redefine statistical significance" printed in *The Political Methodologist*.

3. Daniel Läkens, Assistant Professor in Applied Cognitive Psychology at Eindhoven University of Technology and an author or co-author on many articles on statistical inference in the social sciences, including the Open Science Collaboration's recent *Science* publication "Estimating the reproducibility of psychological science" (Open Science Collaboration, 2015). Daniel is lead author of the pre-print "Justify Your Alpha" (Läkens et al., 2017), a response to "Redefine statistical significance."

4. Blakeley B. McShane, Associate Professor of Marketing at the Kellogg School of Management at Northwestern University and a co-author of the recent paper "Abandon Statistical Significance" (McShane et al., 2017) as well as many other articles on statistical inference and replicability.

5. Jennifer L. Tackett, Associate Professor of Psychology at Northwestern University and a co-author of the recent paper “Abandon Statistical Significance” (McShane et al., 2017) who specializes in childhood and adolescent psychopathology.

6. E.J. Wagenmakers, Professor at the Methodology Unit of the Department of Psychology at the University of Amsterdam, a co-author of the paper in *Nature Human Behavior* (Benjamin et al., 2017) and author or co-author of many other articles concerning statistical inference in the social sciences, including a meta-analysis of the “power pose” effect (Gronau et al., 2017).

Questions

Would the panelists advocate for the presentation of both NHST [null hypothesis significance testing] and Bayes [factors] together to quantify evidence?

–Jason Geller, jgeller1@uab.edu

Daniel Benjamin: The appropriate statistical tools depend on the situation. For example, in economics (my own discipline), the prevailing theory sometimes predicts a particular parameter value, and we want to know whether the data can be rationalized by the theory or not. In that case, there is a null hypothesis but not an alternative hypothesis, and NHST can make sense as an approach.

In settings where there is an alternative hypothesis—and science often proceeds better when studies are designed to pit hypotheses that make different predictions against each other—then we will often be interested in the extent to which the data are more consistent with one hypothesis than another. That is a case where the Bayes factor is a useful tool.

The key point of my presentation was that when we do NHST, a p-value of 0.05 is actually fairly weak evidence against the null hypothesis (certainly weaker than most of us have realized)—at most $\approx$ 3:1 evidence against the null. Presenting Bayes factors alongside p-values would certainly help researchers recognize how strong the evidence against the null hypothesis really is, so I think it would be great if researchers did so whenever they could.

But even if researchers don't calculate Bayes factors, I think we need to recognize that a p-value of 0.05 is fairly weak evidence, and results with a p-value of 0.05 should not be treated and interpreted as if the evidence were strong.

Blakeley McShane and Jennifer Tackett: We are agnostic on the statistical approach (e.g., frequentist, Bayesian, Fisherian, machine learning, or hybrid) taken: lots of approaches will work in lots of situations with some being better suited for some and others for others. Most likely individuals develop expertise in one or more approaches and will do best by sticking to these (e.g., conditional on a given dataset and question, Andrew Gelman might have better luck employing a hierarchical model while Rob Tibshirani might employing the Lasso).

Our point is, instead, twofold. First, we should move away from using purely statistical measures of the evidence (whether a p-value, a Bayes Factor, a confidence interval, or what have you). More specifically, we should avoid using binary (or categorical) thresholds based on these purely statistical measures, and in particular avoid lexicographic decision rules in the publication process and in statistical decision making more broadly.

Second, while treating these purely statistical measures as one source of evidence, we should also pay attention to other important considerations such as prior and related evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain (what we call the heretofore "neglected factors"). In publication, editors and reviewers already pay attention to these factors—just generally only after $p < 0.05$. There is no need for this threshold screening, and it creates many wide-ranging problems.

E.J. Wagenmakers: I have been struggling with this question for well over a decade. If the goal is to "quantify evidence", then I do not think that NHST is appropriate; quantifying evidence is simply not what NHST seeks to achieve. However, when the goal is to convince a skeptical audience that the conclusions are robust, and hold regardless of the statistical paradigm that is applied, then yes, presenting both is an excellent idea. In general, it is hard to argue that researchers should present less information—in the interest of transparency, I believe that they should present more information.

While I agree with what is proposed regarding the full, conscious analysis of data and the lower emphasis on p-values and their thresholds, I would like to hear how you propose this should be implemented in something like clinical trials where hard decisions need to be made for clinical implementation? Publications are one thing, but practice is a completely different thing. Thank you very much for this exciting discussion!

–Melanie Ganz, mganz@nru.dk

Justin Esarey: There is a long and detailed literature on statistical decision theory, and I think this literature speaks most directly to your question. Statistical decision theory studies how to integrate statistical information about uncertain quantities (e.g., causal relationships) with formal evaluation of costs and benefits in a utility function in order to produce a rational (viz., expected utility maximizing) decision. I think it would be useful to create a statistical decision theoretic framework to guide repeated and structured decisions made by regulatory authorities like the Food and Drug Administration. Subjective and/or ethically-laden factors (operationalized as features of the utility function) are involved in making these decisions, such as the relative weighting of possible costs and benefits, and decision theory cannot tell you what these ought to be. In my mind, one role of a decision-making body is to specify and justify these choices and show how they lead to the decision. Another role is to ensure that the empirical evidence that goes into the decision is the product of an appropriate scientific process (that concerns about confounding, selection bias, simultaneity, generalizability, and so on have been addressed such that we have confidence in statistical estimates of uncertain quantities).

Of course, we also have to make decisions about publications! Many of these decisions are binary. We decide whether to accept or reject a paper for publication, whether to conclude that a paper's finding is important, whether to publicize the finding in the media, whether to assign a paper on a syllabus... and so on. In a paper with Nathan Danneman (Esarey and Danneman, 2015), I develop some software that can help scientists make decisions like these in a structured way using statistical decision theory. As a bonus, when used to decide whether to reject the null hypothesis under the NHST, our procedure increases the power (i.e., the probability of correctly rejecting a false null hypothesis) at every size (i.e., the probability of incorrectly rejecting a true null hypothesis) in our simulations. You may find this software useful for your decision problems as well.

Blakeley McShane and Jennifer Tackett: Thank you! We tread lightly on this as it moves from the scientific realm into the political. One way an "Abandon Statistical Significance" approach could be brought into the policy arena is the following (this is a near paraphrase and translation of our advice to editors and reviewers and authors for the clinical trial setting). Perhaps there is no reason "hard decisions" need to be made by policy makers—whether based p-value (or other statistical) thresholds or otherwise. Instead, let patients in consultation with their doctors make these hard decisions (indeed, they already do to a large degree!). Policy makers could play a role in making sure the information reported to patients and doctors is accurate. But, it seems possible for patients in consultation with their doctors to weigh, for example, any benefits of a proposed intervention in terms of efficacy with any costs in terms of side effects and financial implications. Not only will/should the weight placed on each of these factors vary by individual, but efficacy benefits and side effect costs will as well. This also underscores the importance of one of our "neglected factors," real world costs and benefits. If researchers were required to examine and report their results in the context of real world costs and benefits, direct applications would be much more easily obtained from published research. It also is consistent with our call to accept uncertainty in and embrace variation in effects (in this example, most notably uncertainty and variation in costs and benefits of the treatment across individual patients).

E.J. Wagenmakers: In Bayesian inference, one can monitor evidence as the data accumulate. This evidence is graded, but if hard decisions are required one can consider utilities and impose a cutoff value. The point of the "Redefine Statistical Significance" paper was to argue that the current $p < 0.05$ threshold is too lenient.

Question for Jennifer Tackett: What does "novelty of findings" as a neglected factor mean exactly?

–Ahmed Khalil, ahmed.arahim.k@gmail.com

Blakeley McShane and Jennifer Tackett: Journals are typically not interested in publishing well-known findings (e.g., the force of gravitational attraction between two objects is inversely proportional to the square of their distance). However, novelty can sometimes be in conflict with rigor and replicability as new findings are necessarily less tested. We think editors and reviewers are much better suited to make this tradeoff than a $p < 0.05$ or any other threshold-based rule. As an editor, I also typically consider novelty in the context of what any given paper adds to the existing literature: what does it tell us that is new?

What happens to Type I errors if you reduce alpha level to 0.005, and are there not specific cases in science concerned with discovery where missing out on potential effects might be more deleterious than discovering false effects? [I am] interested in a discussion of striking the proper balance between Type I and Type II [errors].

–Eiko Fried, eikofried@gmail.com

Daniel Benjamin: Reducing the alpha level to 0.005 mechanically reduces the Type I error rate, and in our paper, we advocate increasing sample sizes to keep the Type II error rate constant. Of course, that may not always be possible, for example, for researchers studying a dataset whose sample size is fixed. In that case, reducing the alpha level to 0.005 would increase the Type II error rate.

This framing in terms of Type I and Type II errors puts us in the world of statistical decision theory, where we need to make a dichotomous decision. In that world, I agree with the “Justify Your Alpha” contingent: we should set the threshold for making a decision to optimally balance the Type I versus Type II error rates, and the optimal threshold depends on the costs of each type of error.

But in scientific research, we’re usually not conducting our study in order to make a particular decision. Instead, we’re often trying to understand how strong the evidence is for or against particular hypotheses in order to update our beliefs. It’s not that we’re going to make some decision that depends on whether the result is “statistically significant,” but rather we’re going to conclude that the evidence strongly disfavors the null hypothesis and update our beliefs accordingly. For that purpose, we need to correctly interpret how strongly a particular p-value disfavors the null. For a p-value of 0.05, the answer is: not much. The way I think about reducing the alpha level to 0.005 is that it allows us to be more correct when we treat a “statistically significant” result as implying strong evidence against the null. But if the p-value is 0.05, we shouldn’t “accept the null”—we should treat the evidence as merely suggestive against the null. (In this way of thinking about p-values, the Type II error rate is relevant for planning the study: we’d like to have sufficient power to get a statistically significant result. But once the study has been conducted, we should update our beliefs on the basis of the actual, observed p-value, not merely whether it is larger or smaller than alpha.)

I agree with the “Abandon Statistical Significance” contingent that it is even better to treat the p-value as a continuous measure of the strength of evidence, and often we would want to supplement or replace the p-value with other statistical measures (depending on the research question). But even if we treat the p-value as a continuous measure, it’s important to correctly interpret what 0.05 (or any other value) means in terms of the strength of evidence against the null.

Justin Esarey: I think everyone agrees that striking a proper balance between Type I and Type II errors is important. Indeed, making researchers think explicitly about this balance when choosing to draw an inference is something that Nathan Danneman and I (Esarey and Danneman, 2015) incorporated into our statistical decision procedure.

I think it's important to remember that lowering α to 0.005 probably won't have the simple effect of lowering power that we would expect in a single statistical model on a fixed data set, where the tradeoff between Type I and Type II errors is direct. This is so because researchers can take actions to lower Type I error without increasing Type II error. Some are simple, but can be costly or even impossible in some scenarios; one example of such a measure is increasing the sample size for a given analysis. For this reason, lowering α to 0.005 may encourage larger samples (thus improving the replicability of research without harming the power of a study to make new discoveries), but may also lower the overall productivity of scientists (as researchers must devote more resources to collecting larger samples in fewer projects). It may also discourage researchers from doing studies where this is impossible, making it harder to make discoveries that are only possible in fixed-N data sets.

Another measure to lower Type I error (and the one that I discuss in my article in *The Political Methodologist*) is to pre-specify a larger number of different hypotheses from a theory and to jointly test these hypotheses. Because the probability of simultaneously confirming multiple disparate predictions by chance is (almost always) lower than the probability of singly confirming one of them, the size of each individual test can be larger than the overall size of the test, allowing for the possibility that the overall test is substantially more powerful at a given size.

The upshot is that any reform, including lowering α to 0.005, will have complex effects on the research ecosystem (including the prevalence of Type I and II errors) that are hard to know in advance. For this reason, experimentation with this and other reforms in different journals is something I've begun to see as appealing.

Blakeley McShane and Jennifer Tackett: Indeed. This will vary wildly by domain—hence our belief that that p-value thresholds (as well as those based on other statistical measures) are a bad idea in general.

However, we add that we dislike the Type I/Type II error dichotomous framework in the biomedical and social sciences. We believe various features of contemporary biomedical and social sciences—small and variable effects, noisy measurements, a publication process that screens for statistical significance, and research practices—make NHST and in particular the sharp point null hypothesis of zero effect and zero systematic error particularly poorly suited for these domains. Indeed, the sharp point null hypothesis of zero effect and zero systematic error used in the overwhelming majority of applications in the biomedical and social sciences is generally not of interest because it is generally implausible. Thus, the Type I/Type II error framework makes little sense in these domains.

Is this debate worth having, when it is apparent (to me, anyway) that the largest undermining force in these sciences is not the threshold, but rather the incentive structures that encourage people to dive over these thresholds, cherry pick results, engage in QRPs [questionable research practices], etc? I think a big fear about the 0.005 paper was simply that it would itself undermine efforts to change these incentive structures.

–Stephen Martin, stephen_martin@baylor.edu

Daniel Benjamin: I agree that the threshold is not the biggest problem, but I think it is an underappreciated contributing factor.

If your fear is justified—if the 0.005 proposal in fact undermined efforts to address the bigger problems—then I would stop advocating for 0.005. But I don't think the 0.005 proposal will undermine other efforts. As we say in the paper, we intend the proposal to be a complement to other reforms, not a substitute for them. And I haven't seen anyone arguing that changing the threshold to 0.005 is all we should do.

Blakeley McShane and Jennifer Tackett: Well said. To paraphrase Mickey Inzlicht, the academic incentive structure is fucked—and the task of revamping it is undoubtedly much more challenging than the statistical challenges we are discussing here.

E.J. Wagenmakers: I agree that the current incentive structure is toxic. However, we should still demand compelling evidence for bold empirical claims (e.g., "we reject the null hypothesis"), and this means we must abandon the $p < 0.05$ threshold. One might speculate that lenient thresholds invite statistical abuse.

I'd love to hear/see examples of papers that have been used taking the "abandon statistical significance" approach of treating p-values without thresholds per se.

–Sam Parsons, sam.parsons@psy.ox.ac.uk

Blakeley McShane and Jennifer Tackett: Not to be impertinent, but see much work by the authors of the "Abandon Statistical Significance" paper: Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer Tackett. This approach is often in evidence there. In addition, there are also certain disciplines and research areas (e.g., work on psychopathology structure using structural equation modeling) that seldom if ever employ $p < 0.05$ thresholds in interpreting results and drawing conclusions. The general concept is thus not new at least in some areas of social science research.

To Daniel Benjamin: Which current academics were most influential in your own reasoning for [lowering the threshold of statistical significance to 0.005]?

–Sameera Daniels, sierra_bardot@yahoo.com

Daniel Benjamin: For me, an "aha" moment was reading Sellke, Bayarri, and Berger (2001). They provided a simple formula, which holds under quite general conditions, for the upper bound on the Bayes factor implied by any given p-value. They pointed out that a p-value of 0.05 corresponds to at best only ≈ 3:1 evidence against the null hypothesis. Other researchers have drawn the same conclusion under different assumptions (e.g., Edwards, Lindman, and Savage (1963); Johnson (2013)), but I first came across the conclusion from reading Sellke, Bayarri, and Berger.

The medical genomics research community also deeply influenced me. Much of my research is genomics applied to behavioral traits, so I've become familiar with the traditions in medical genomics. In the mid-2000s, there was a crisis of confidence in the research findings in medical genomics, much like the "replication crisis" that has beset the social sciences in the past few years. Within a few years, that community transitioned to a new set of standard research practices, including much larger sample sizes and a p-value threshold of 5×10^{-8}. The main argument for adopting that threshold was to correctly adjust for multiple hypothesis testing in genome-wide studies, which is different than the argument my co-authors and I are making for 0.005 in the social sciences. But the fact that the community was able to quickly adapt to the new threshold—and that doing so contributed importantly to making research results much more robust and replicable—played a major role in inspiring me to advocate for a change in the significance threshold.

It's important to note that by the time I realized 0.005 was a good idea, Val Johnson had already been advocating it for several years.

Could one of the presenters discuss explicitly some of the trade-offs between power and other sources of error? It seems, for example, like lowering standard alpha levels would increase the number of studies with large sample sizes, but decrease the number of studies with good designs (because good designs typically cost more, and sometimes have hard upper limits on sample sizes)?

–Anonymous Attendee

Daniel Benjamin: The relevant tradeoffs are going to depend on the research topic. It's worth point out, though, that there are already many such tradeoffs. For example, it's often easier to get large, representative samples for surveys than for experiments. These tradeoffs generate an opportunity set of possible research studies that we could undertake. The optimal mix of studies to conduct depends on our assessment of the relative importance of the different dimensions of the tradeoff. I believe that one of the lessons of the replication crisis is that we haven't been assigning enough importance to sample size. If we start doing so, then it will shift the mix of optimal studies toward those with larger sample sizes, at the cost of other dimensions of good research.

Blakeley McShane and Jennifer Tackett: This is unclear. Good designs often mean better, more precise measures. Better, more precise measures allow for the same power with smaller sample sizes. Nonetheless, we do not like to think in terms of power, which is inherently bound up with the NHST framework. Instead, we'd also like to see researchers take a more direct route to good science via more careful theorizing, more precise individual-level measurements, a greater use of within-person or longitudinal designs, and increased consideration of models that use informative priors, that feature varying treatment effects, and that are multilevel or meta-analytic in nature.

I would love to hear the pro-0.005 people talk about how 0.005 would work within the current scientific publishing incentive structure. Given we see p-hacking and HARKing [hypothesizing after the results are known] and in (hopefully) rare instances outright fraud at "only" 0.05, I can definitely see a case for the argument that without a complete overhaul of the system (which people are trying to do, of course), 0.005 will only make those problems worse. Ditto for journals using 0.005 as a cutoff point for publication, which I know the Benjamin et al paper explicitly said they aren't suggesting. (It doesn't mean they won't.)

–Anonymous Attendee

Daniel Benjamin: Changing the significance threshold to 0.005 won't solve all the problems, and it won't even solve the biggest problems, such as p-hacking and HARKing. But I think it will help with one of the problems, which is that people think 0.05 corresponds to stronger evidence against the null hypothesis than it really does. If researchers reserve the term "significant" for p-values below 0.005 and use the term "suggestive" for p-values between 0.05 and 0.005, it will be an important step toward more accurate interpretation and communication of results.

It's conceivable that 0.005 could make things worse, for example, if researchers p-hack to the new threshold. But researchers might also be constrained in the amount of p-hacking they do: a little bit of p-hacking can be justified by realizing that the analyses would be better done a different way, but a lot of p-hacking feels like cheating, and most researchers don't want to knowingly cheat. One benefit of the 0.005 threshold is that it makes it harder to get a significant result for any given amount of p-hacking. In any case, even after changing the threshold to 0.005, we need reforms to minimize the amount of p-hacking, HARKing, etc., such as incentivizing preregistration and transparent reporting of all analyses conducted.

Justin Esarey: I mentioned this point in a previous response, but it bears repeating: any reform, including lowering α to 0.005, will have complex effects on the research ecosystem that are hard to know in advance. I think the most honest answer to your question that I can give is that no one's sure exactly what effect this change would have on the overall population of published research, just as Fisher and Neyman and Pearson could not have known all the effects that the NHST with $\alpha = 0.05$ would have on applied statistical work. When we don't know the answer to an empirical question concerning social behavior, we typically initiate an empirical research program to determine the answer. I don't see why we can't do the same here!

E.J. Wagenmakers: In my opinion, it will be a lot more difficult to p-hack your way to 0.005 than to 0.05. With a lenient threshold, subtle measures suffice to get p below 0.05, and researchers may hardly notice that they are using questionable research practices. But when you have to pass the 0.005 threshold, you really have to torture the data; my hope is that many researchers will not have the stomach for this. Consequently, researchers may develop more of a tolerance for null results. But such speculations are beside the point: we should simply not accept that bold claims are based on weak evidence.

What will happen when reviewer 1 tells me to remove the p-values, reviewer 2 rejects because $p \not< 0.005$*, and reviewer 3 asks for justification? Will the editor make the final decision?*

–Anonymous Attendee

Blakeley McShane and Jennifer Tackett: Yes—just like the editor does now when reviewer one loves the paper, reviewer two hates it, and reviewer three is lukewarm. Or when reviewer one thinks the paper is strong on theory but weak on empirics, reviewer two thinks it strong on empirics but is dissatisfied with the data/experimental design, and reviewer three thinks it competently done but lacks real world/policy/managerial implications.

Rather than changing thresholds or other large scale changes, should we first focus on improving stats teaching?

–Sam Parsons, sam.parsons@psy.ox.ac.uk

Daniel Benjamin: We should certainly focus on improving stats teaching, but the payoffs from doing so will be long-term. I don't view the process of reform as either-or: there are many things we should be doing all at once. This includes incentivizing preregistration and transparent reporting of all analyses conducted and changing norms away from threshold thinking, all while also improving stats teaching to incorporate these reforms. Changing the significance threshold won't solve the bigger problems with statistical practice, but it will have immediate, positive benefits in terms of researchers interpreting and communicating the strength of evidence more accurately.

Blakeley McShane and Jennifer Tackett: No disagreement from any of us on this one. As the recent American Statistical Association Statement on Statistical Significance and p-values (Wasserstein and Lazar, 2016) stated:

> In February, 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

> Q: Why do so many colleges and grad schools teach p = .05?
>
> A: Because that's still what the scientific community and journal editors use.
>
> Q: Why do so many people still use p = 0.05?
>
> A: Because that's what they were taught in college or grad school.
>
> Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: "We teach it because it's what we do; we do it because it's what we teach."

Similarly, as our coauthor Andrew Gelman wrote in his comment on this Statement (Gelman, 2016):

> I put much of the blame on statistical education, for two reasons.
>
> First, in our courses and textbooks (my own included), we tend to take the "dataset" and even the statistical model as given, reducing statistics to a mathematical or computational problem of inference and encouraging students and practitioners to think of their data as given. Even when we discuss the design of surveys and experiments, we typically focus on the choice of sample size, not on the importance of valid and reliable measurements. The result is often an attitude that any measurement will do, and a blind quest for statistical significance.
>
> Second, it seems to me that statistics is often sold as a sort of alchemy that transmutes randomness into certainty, an "uncertainty laundering" that begins with data and concludes with success as measured by statistical significance. Again, I do not exempt my own books from this criticism: we present neatly packaged analyses with clear conclusions. This is what is expected—demanded—of subject-matter journals. Just try publishing a result with p = 0.20. If researchers have been trained with the expectation that they will get statistical significance if they work hard and play by the rules, if granting agencies demand power analyses in which researchers must claim 80% certainty that they will attain statistical

> significance, and if that threshold is required for publication, it is no surprise that researchers will routinely satisfy this criterion, and publish, and publish, and publish, even in the absence of any real effects, or in the context of effects that are so variable as to be undetectable in the studies that are being conducted (Gelman and Carlin 2014).

Also, as one of us (McShane along with Gal) has written elsewhere (McShane and Gal, 2017):

> [S]tatistics at the undergraduate level as well as at the graduate level in applied fields is often taught in a rote and recipe-like manner that typically focuses nearly exclusively on the NHST paradigm. To be fair, statisticians are only partially at fault for this: statisticians are often not responsible for teaching statistics courses in applied fields (this is probably especially the case at the graduate level as compared to the undergraduate level) and, even when they are, institutional realities often constrain the curriculum.
>
> The recent trend toward so-called "data science" curricula may prove helpful in facilitating a reevaluation and relaxation of these institutional constraints. In particular, it may provide statisticians with the institutional leverage necessary to move curricula away from the rote and recipe-like application of NHST in training and toward such topics as estimation, variability, and uncertainty as well as exploratory and graphical data analysis, model checking and improvement, and prediction. Further, these curricula may help facilitate a move away from point-and-click statistical software and toward scripting languages. This in and of itself is likely to encourage a more holistic view of the evidence; for example, data cleaning in a scripting language naturally prompts questions about the quality of the data and measurement while coding a model oneself increases understanding and likely promotes deeper reflection on model specification and model fit. Thus, recent developments in curricula may well help mitigate dichotomous thinking errors.

Regarding significance and publication bias: What is your expectation about the impact of setting $\alpha = 0.005$ *on publication bias if journals (by that I mean editors and reviewers less involved in the current discussion) were to adapt this thinking without changing their selection? And extending on that: If we abandon the significance criterion, isn't there a risk that journals will introduce something similar implicitly, e.g. by checking if the credibility intervals include 0?*

–Christopher Harms, christopher.harms@uni-bonn.de

Justin Esarey: I directly address the question about publication bias in my paper for *The Political Methodologist* using simulation analysis; see especially Figure 3 in that paper. For any single relationship being studied, publication bias will increase when α is increased to 0.005 if journals continue to accept only statistically significant results using the new threshold. However, when considering the overall level of publication bias in a population, the level of publication bias can actually *decline* due to a stricter α. This can occur because the number of published results corresponding to true null hypotheses falls. Therefore, the average gap between the true effect and the estimated (published) effect—which is often particularly large when the null hypothesis is true—declines.

Blakeley McShane and Jennifer Tackett: Responding to the extension, we are very clear to rule this out:

> What can be done? Statistics is hard, especially when effects are small and variable and measurements are noisy as in the biomedical and social sciences. There are no quick fixes. Proposals such as changing the default p-value threshold for statistical significance, employing confidence intervals with a focus on whether or not they contain zero, or employing Bayes factors along with conventional classifications for evaluating the strength of evidence suffer from the same or similar issues as the current use of p-values with the 0.05 threshold. In particular, each implicitly or explicitly categorizes evidence based on thresholds relative to the generally uninteresting and implausible null hypothesis of zero effect and zero systematic error. Further, each is a purely statistical measure that fails to

take a more holistic view of the evidence that includes the consideration of the traditionally neglected factors, that is, prior and related evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain.

In brief, each is a form of statistical alchemy that falsely promises to transmute randomness into certainty, an "uncertainty laundering" (Gelman, 2016) that begins with data and concludes with dichotomous declarations of truth or falsity—binary statements about there being "an effect" or "no effect"—based on some p-value or other statistical threshold being surpassed. A critical first step forward is to begin accepting uncertainty and embracing variation in effects (Carlin, 2016; Gelman, 2016) and recognizing that we can learn much (indeed, more) about the world by forsaking the false promise of certainty offered by such dichotomization.

Bayes factor people deny the relevance of error probabilities. Therefore they're unable to pick up on biasing selection effects such as optional stopping, cherry-picking, and post-data subgroups. This goes diametrically against the goals of preregistration and adjusting for multiple testing. So there's a serious tension with adopting a Bayes account.

–Deborah Mayo, error@vt.edu

E.J. Wagenmakers: This is an interesting issue. I am a Bayes factor person and I also advocate preregistration and corrections for multiple testing. In the Bayesian world, cherry-picked hypotheses are relatively implausible a priori, and cherry-picked analysis pipelines affect the likelihood (as they do in Bayesian models for publication bias). So there is no tension.

If we are calling for methodological plurality in journal approaches, then how does one help promote that? It seems that hasn't changed much despite attempts in a variety of journals to promote pre-acceptance. One of the problems with having variety is that you would have to rewrite your results for each journal type which would be a huge problem for researchers with limited time.

–Aaron Erlich, aaron.erlich@mcgill.ca

Justin Esarey: In many cases, a journal's standards for quantitative evidence are determined by its editor(s), and to a lesser extent by its editorial board and reviewer pool. Thus, editors should be largely free to impose the policies that they think will be most beneficial—and I expect that these policies will be diverse, if they reflect the diversity of opinion in the wider pool of researchers. In some cases, significant policy revisions have already been made (e.g., the editor of *Basic and Applied Social Psychology* does not allow use of the NHST in work published there).

I think that variation in journal policies would be more scientifically valuable as an exploration of the effect of publication standards on research if it were more systematic and structured, not implemented via the independent and uncoordinated decisions of many editors. However, I acknowledge that coordinating such a project would be a significant and difficult undertaking comparable to, and perhaps even more challenging than, the Many Labs replication project.

E.J. Wagenmakers: I agree this is a problem, and this is why our group at the University of Amsterdam has developed JASP, a free and open-source program that provides both frequentist and Bayesian methods (`https://www.jasp-stats.org`). With JASP, a Bayesian reanalysis is often a matter of executing two or three mouse clicks.

How would use of some of the Bayesian methods proposed (assuming they become more widely used which would imply at least, incorporation into standard statistical packages) affect meta analyses; can they be easily incorporated with the use of more traditional p-value studies?

–Brian Finch, brian.finch@usc.edu

Blakeley McShane and Jennifer Tackett: Meta-analytic models are inherently hierarchical / multi-level in nature. Such models, even if estimated using classical techniques like REML, inherently possess a somewhat Bayesian flavor. A fully Bayesian treatment would just make the priors and hyperpriors more explicit. And certainly the way a primary study is analyzed has no/little impact on how it might be incorporated into a meta-analysis; what matters is how that study reports the data and choices made by the meta-analyst.

How do you propose dealing with the fact that $p < 0.05$ *is written up in laws and policies within the government as a standard of evidence?*

–Laura Kapitula, kapitull@gvsu.edu

Blakeley McShane and Jennifer Tackett: This is a sad reality. We can only hope (perhaps in vain) that the government will catch up with decades old advances in statistics.

E.J. Wagenmakers: Throughout the ages, most governments have at one point or another promoted the most damaging, degrading, and outright ridiculous laws: witch-burning, slavery, male-only voting, limited gun-control. Most modern nations have learned from their mistake and abandoned those laws. But some senseless laws persist.

Here we have a group of experts, who really enjoy diving into statistical questions. It's great to watch the discussion. However: If I think of the average "user" of statistical methods in my department, really most of them have no clue of ***any*** *of these positions, let alone be able to "justify their own alpha" (they would always end up at 5%, or maybe go to 10%, which is a one-sided test...), "judge the continuous holistic pattern of evidence", etc. And they are much too busy writing grants, papers, etc., to find the time to read all of these great papers. My question to each of the panelists: Do you think that your approach can be* ***realistically*** *implemented for typical social science researchers?*

–Anonymous Attendee

Justin Esarey: I think that most of the proposed alternatives that we have discussed don't require a substantially greater depth of statistical sophistication to implement than the NHST. In the case of lowering α to 0.005, the procedure is the same as before but with a different target p-value. Using Bayes' factors or Bayesian credible intervals explicitly in the place of p-values wouldn't be that much more complicated, and indeed I guess that the interpretation of these statistics would comport better with people's intuitions about probability compared to frequentist reasoning. The proposal to "justify your α" requires a reasonable understanding of how and why the NHST works the way it does, but it's still build around the same inferential process. To explicitly use statistical decision theory in the presentation and adjudication of results might be a little more complicated, but no more complicated than estimating a maximum likelihood model—and, as when we estimate a maximum likelihood model, well-designed software can make the process easier for everyone to implement.

Blakeley McShane and Jennifer Tackett: Current researchers? Perhaps. Perhaps not. If not, perhaps what we need is smarter and better-trained—and thus perhaps inevitably fewer—researchers. Indeed, fewer researchers and thus less research output could be a good thing given current widespread replication failure (under the mild proviso that smarter and better-trained researchers are more likely to produce more reliable results). A move toward slower science could benefit us all. Of course, there is a difficult balance vis-à-vis the incentive structure, as noted in response to a previous question. But ultimately, we should all be more concerned with producing meaningful, reproducible work than falling back on the flawed simplistic heuristic of churning out as many papers as quickly as possible.

References

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E. J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcom Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, James Holland Jones, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson. 2017. "Redefine Statistical Significance." *Nature Human Behavior* Forthcoming:1–18. URL: `https://www.nature.com/articles/s41562-017-0189-z` accessed 12/3/2017.

Carlin, John B. 2016. "Is reform possible without a paradigm shift?" *The American Statistician (supplemental material)* . URL: `https://ndownloader.figshare.com/files/5368454` accessed 12/3/2017.

Edwards, Ward, Harold Lindman and Leonard J Savage. 1963. "Bayesian statistical inference for psychological research." *Psychological review* 70(3):193.

Esarey, Justin. 2017. "Lowering the Threshold of Statistical Significance to $p < 0.005$ to Encourage Enriched Theories of Politics." *The Political Methodologist* 24(2):13–20. URL: `https://thepoliticalmethodologist.files.wordpress.com/2017/09/v24-n2-fix.pdf`.

Esarey, Justin and Nathan Danneman. 2015. "A Quantitative Method for Substantive Robustness Assessment." *Political Science Research and Methods* 3(1):95111. URL: `http://dx.doi.org/10.1017/psrm.2014.14`.

Gelman, Andrew. 2016. "The problems with p-values are not just with p-values." *The American Statistician (supplemental material)* . URL: `https://ndownloader.figshare.com/files/5368460` accessed 12/3/2017.

Gronau, Quentin F., Sara Van Erp, Daniel W. Heck, Joseph Cesario, Kai J. Jonas and Eric-Jan Wagenmakers. 2017. “A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: the case of felt power.” *Comprehensive Results in Social Psychology* 2(1):123–138. URL: `https://doi.org/10.1080/23743603.2017.1326760`.

Johnson, Valen E. 2013. “Revised standards for statistical evidence.” *Proceedings of the National Academy of Sciences* 110(48):19313–19317.

Läkens, Daniel, Federico G. Adolfi, Casper Albers, Farid Anvari, Matthew A.J. Apps, Shlomo Engelson Argamon, Marcel A.L.M. van Assen, Thom Baguley, Raymond Becker, Stephen D. Benning et al. 2017. “Justify Your Alpha: A Response to ‘Redefine Statistical Significance’.” Online (version 9/18/2017). URL: `https://psyarxiv.com/9s3y6` accessed 12/3/2017.

McShane, Blakeley B. and David Gal. 2017. “Statistical Significance and the Dichotomization of Evidence.” *Journal of the American Statistical Association* 112(519):885–895.
URL: *https://doi.org/10.1080/01621459.2017.1289846*

McShane, Blakeley B., David Gal, Andrew Gelman, Robert Christian and Jennifer L. Tackett. 2017. “Abandon Statistical Significance.” Online (v. 21 September 2017). URL: `http://www.stat.columbia.edu/~gelman/research/unpublished/abandon.pdf`.

Open Science Collaboration. 2015. “Estimating the reproducibility of psychological science.” *Science* 349(6251):aac4716. URL: `http://science.sciencemag.org/content/349/6251/aac4716` accessed 8/5/2017.

Sellke, Thomas, M.J. Bayarri and James O. Berger. 2001. “Calibration of ρ values for testing precise null hypotheses.” *The American Statistician* 55(1):62–71.

Wasserstein, Ronald L. and Nicole A. Lazar. 2016. “The ASA’s Statement on p-Values: Context, Process, and Purpose.” *The American Statistician* 70(2):129–133. URL: `http://dx.doi.org/10.1080/00031305.2016.1154108` accessed 8/5/2017.