IRTP INTERNATIONAL REVIEW OF THEORETICAL PSYCHOLOGIES

• Vol. 1, No. 2 • 2021 • (pp. 178-192) •
www.istp-irtp.com

DOI: https://doi.org/10.7146/irtp.v1i2.127764

Perhaps Psychology's Replication Crisis is a Theoretical Crisis that is Only Masquerading as a Statistical One

Christopher D. Green

Department of Psychology, York University

Abstract

The "replication crisis" may well be the single most important challenge facing empirical psychological research today. It appears that highly trained scientists, often without understanding the potentially dire long-term implications, have been mishandling standard statistical procedures in the service of attaining statistical "significance." Exacerbating the problem, most academic journals do not publish research that has not produced a "significant" result. This toxic combination has resulted in journals apparently publishing many Type I errors and declining to publish many true failures to reject H_0. In response, there has been an urgent call from some psychologists that studies be registered in advance so that their rationales, hypotheses, variables, sample sizes, and statistical analyses are recorded in advance, leaving less room for post hoc manipulation. In this chapter, I argue that this "open science" approach, though laudable, will prove insufficient because the null hypothesis significance test (NHST) is a poor criterion for scientific truth, even when it is handled correctly. The root of the problem is that, whatever statistical problems psychology may have, the discipline never developed the theoretical maturity required. For decades it has been satisfied testing weak theories that predict, at best, only the direction of the effect, rather than the size of effect. Indeed, uncritical acceptance of NHST by the discipline may have served to stunt psychology's theoretical growth by giving researchers a way of building a successful career without having to develop models that make precise predictions. Improving our statistical "hygiene" would be a good thing, to be sure, but it is unlikely to resolve psychology's growing credibility problem until our theoretical practices mature considerably.

Keywords: Meehl, psychology, replication, statistics, theory

For many psychologists, the "replication crisis" in psychology began in 2011, in the weeks immediately after it was made known on social media that the prestigious *Journal of Personality and Social Psychology* had accepted for publication an article by Daryl Bem

(2011) that claimed to provide substantial new evidence for precognition. A preprint of the article was widely circulated on the internet and, soon, trenchant critiques as well as failed replication efforts started to appear. Many of these were apparently submitted for publication to *JPSP*, but the journal's response was, as per their policy, that they publish only "original" research: neither critiques nor replications. The outrage over this stance was widespread. That one of psychology's leading journals was (1) not only going to publish psychical research, which of its very nature was highly dubious, but also (2) that the journal would not allow members of the discipline to respond to this singular event was certain to undermine whatever scientific reputation experimental psychology had managed to build for itself over the previous decades.

Eventually, the journal allowed a single reply, a critique by the Dutch mathematical psychologist E. J. Wagenmakers and three colleagues, pointedly titled "Why psychologists must change the way they analyze their data" (Wagenmakers et al., 2011). Rather than focusing on the details of psychical research specifically, Wagenmakers and his colleagues used their unique opportunity to call into question those conventional statistical practices of the discipline that enabled Bem to claim evidence for precognition that was so *apparently* compelling that *JPSP*'s reviewers and editors felt duty-bound to publish this highly debatable research. The problem, Wagenmakers argued, was not to be found in particular method Bem had used in investigating his topic but, much more ominously, in the heavy reliance that nearly all of psychology, including Bem, conventionally placed on the adequacy of null hypothesis significance testing (NHST) to identify the truth. Wagenmakers had been a critic of NHST for several years by that point (see Wagenmakers, 2007), promoting the Bayesian approach to statistical analysis instead. But, like a number of Bayesians before him (Bakan, 1953, 1966; Edwards et al., 1963; Rozeboom, 1960), he had not produced much visible impact on the discipline. The Bem precognition fiasco, however, focused the attention of the discipline on the weakness on psychology's statistical traditions in a way it had not been since the American Psychological Association's blue ribbon "Task Force on Statistical Inference" had released its landmark report more than a decade earlier (Wilkinson, 1999).

Wagenmakers' criticism of NHST and, by extension, of Bem's precognition research, was that the conventional $p < .05$ criterion for "significance" is far too weak to serve as a basis for serious scientific inference. He went on to show that if Bem's data were reanalyzed using basic Bayesian procedures, it could be easily shown that the evidence provided, though it mostly passed the conventional $p < .05$ criterion, was actually quite weak according to Bayesian standards; certainly not adequate to force a monumental overturning of scientific consensus that would be required for us to accept precognition as fact. In short, the problem was not Bem's long-standing fondness for the psychical. The problem was psychology's even longer-standing fondness for a quite weak form of statistical analysis.[1]

[1] Bem has since buttressed his findings with a meta-analysis (Bem et al., 2015). However, meta-analysis is implicated in the replication crisis as much as any statistical procedure. As U. Oregon Sanjay Srivastava once put the matter on Twitter (7 Dec 2018), conducting a meta-analysis of only published results is like trying to estimate the average height of high school students by measuring the heights of the varsity basketball team. In addition, Bem's case was probably not helped by his later comment to a journalist: "If you looked at all my past experiments, they were always rhetorical devices. I gathered data to show how my point would be made. I used data as a point of persuasion, and I never really worried about, 'Will this replicate or will this not?'" (Engber, 2017).

Nearly immediately, the search was on for other areas of psychology in which apparently "significant" effects might not stand up to scrutiny. In 2012, Brian Nosek gathered together 269 co-investigators in a colossal effort named the Reproducibility Project, the aim of which was to attempt the replication of 100 psychological phenomena that were presumably well-established in the scientific literature. As is now well-known, Nosek and his team found that less that 40% of those findings could be replicated, even according to the minimal criterion of attaining a statistically "significant" outcome. They also found that, even among those phenomena that successfully crossed this rudimentary threshold, many showed markedly smaller effects in the replications than in the original studies (Open Science Collaboration, 2015).

The publication of the Reproducibility Project article in what is perhaps America's most prestigious scientific journal, *Science*, provoked an explosive controversy that continues to this day. Many of those whose most influential work had failed to replicate immediately got to work, not shoring up their own research but, rather, attacking the competence and even the motives of those who had conducted the replications. Probably the strongest of the criticisms leveled against the Reproducibility Project was that there was no reason to prefer the outcome of a single negative replication to the outcome of the original study. Indeed, the phenomena of the original studies would seem to have been already replicated in subsequent published studies, which it why they had seemed "well-established" to Nosek and his team in the first place. Then again, there was no way of knowing how many times other researchers had attempted and failed to replicate those effects, then sequestered their findings, unpublished, in the proverbial "file drawer" (Rosenthal, 1979).

It is interesting to compare the findings of the Reproducibility Project to the results of Geoff Cumming's (2009) demonstration (using his ESCI simulation software) that, given two samples of n=32 and a population effect size of d=0.5 (both of which are a little better than the typical published psychological study), one would not expect much more than half of independent-samples t-tests to come up significant.[2] Although the debate over Nosek's replication results has been dominated by questions of p-hacking, fraud, "hidden moderators," incompetent replicators, and the like, the sheer mechanics of NHST, even when computed and applied correctly, could explain the bulk of the failed replications.

In addition, at least as interesting as Cumming's simulation is Alex Etz's (2015) re-analysis of the Reproducibility Project's results, in which NHST was replaced by Bayes Factors. Etz showed that, while about 20% of Nosek's replications were clear failures and perhaps 25% were clear successes, under accepted Bayesian criteria for the strength of evidence (Jeffreys, 1961), more than half of the replication attempts did not produce results strong enough to determine what had happened, one way or the other.

The only solution to these questions was to attempt replications with much greater statistical power than that possessed by either the original studies or the attempted replications. To this end, Nosek assembled three new, even vaster teams of researchers to attempt to replicate key phenomena *in numerous independent labs*, using much larger samples than in the original studies. Across three of these "Many Labs" projects (Ebersole & 63 others, 2016; Klein & 50 others, 2014; Klein & 173 others, 2018), using more than 24,000 participants at more than 180 distinct sites, the success rate was slightly above half. This was higher than

[2] The power of each test is actually .52, the exact probability of obtaining significant results if the null hypothesis were false, under these conditions.

the original 40% of the Reproducibility Project, but hardly a figure inspiring great confidence in the robustness of psychology's published findings.

Even before Nosek's replication research began in 2012, a number of articles had begun to suggest that something had gone terribly awry with the conventional statistical practices employed by many psychologists. Masicampo and Lalande (2012), for instance, scraped and plotted more than 3600 *p*-values from 3 major journals (12 issues each of *Journal of Experimental Psychology: General*, *Journal of Personality and Social Psychology*, and *Psychological Science*). The frequencies of those *p*-values lie quite close to the curve that should theoretically describe them (Sellke et al., 2001), except for the one indicated by the red arrow (which I have added to the original figure). That is the frequency of *p*-values just below the critical value of *p*=.05, sitting at about *double* its expected frequency. It appears that researchers have been doing something to "nudge" their *p*-values from just above .05 back over "the line" to just below .05, and they have been doing it regularly enough that it shows up clearly on a simple frequency plot.

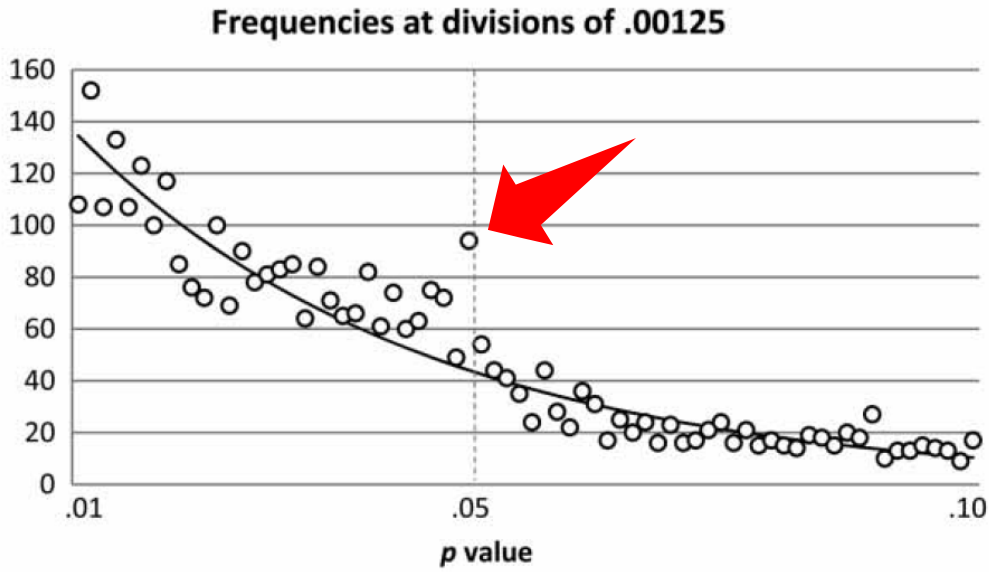

Figure 1. Frequencies of p-values in Masicampo & Lalande (2012). Added red arrow indicates greatly elevated frequency of p-value just below conventional .05 threshold.

Daniel Lakens' (2014) has criticized Masicampo and Lalande's underlying assumption that the *p*-values in published research should lie along so smooth a curve. He argued that, due to publication bias in favor of "significant" findings, we should expect there to be a sharp drop in the frequencies of *p*-values just above the .05 threshold. This is certainly true, and it would explain why *all* of the frequencies between .03 and 05 are above the expected curve, while the great majority of the frequencies between .05 and .07 are below the expected curve. However, this line of thought provides no explanation for why the frequency immediately below .05 (indicated by the red arrow) is so much higher than its neighbors.

Whatever the truth of Lakens' critique, Masicampo & Lalande are hardly alone in having detected substantial anomalies in the frequency of *p*-values just below $p = .05$, and not only in psychology. De Winter and Dodou (2015) surveyed a wide range of scientific literatures and found that the proportion of articles reporting *p*-values between .041 and .049 has increased approximately 10-fold since 1990. Later, Johns Hopkins biostatistician Jeffrey

Leek (2017) plotted 2.5 million *p*-values across the literatures of 25 disciplines. In every discipline, there was an anomalous frequency peak at $p = .05$ (see Figure 2). If you look closely, you will often see another little bump at $p = .10$. It appears that some people may be *p*-hacking merely to be able write the woeful words, "marginally significant."

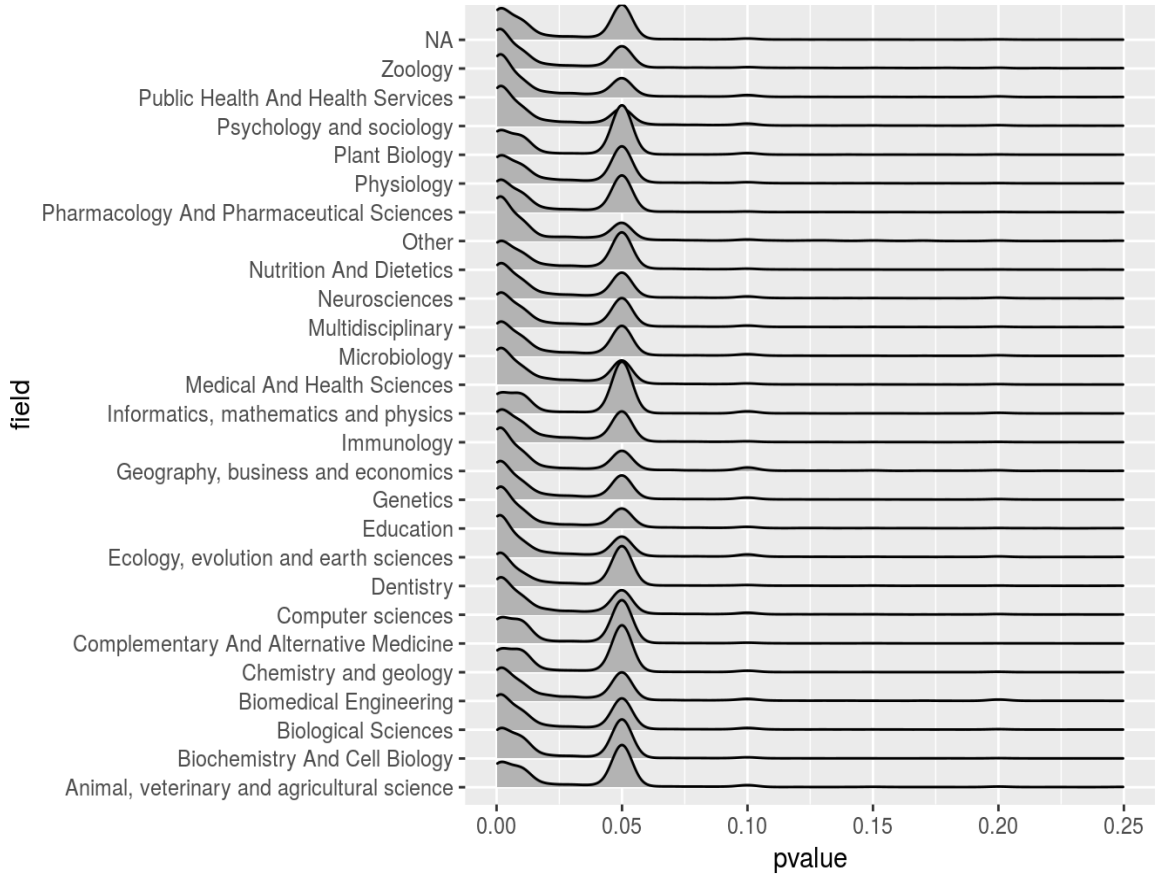

Figure 2. Jeffrey Leek's (2017) plot of the frequencies of a range p-values across 25 disciplines, re-plotted by Nick Fox (personal communication, August 2017) to emphasize the portion of the abscissa below p = .20. Every discipline but economics showed a substantial and unexplained "bump" near p=.05. Note also that frequently there are minor "bumps" near p=.10.

These sorts of anomalies are much more troubling than the simple failures of replication seen before (because they are unlikely to have resulted from of the inadequacies of NHST as an analytic tool). What might be causing these anomalous distributions of *p*-values to appear? A number of possibilities were outlined in the 2011 article titled "False-positive psychology" by Simmons, Nelson, and Simonsohn (2011). They ran simulations of the growth of Type I error rates when initially a researcher runs a small sample, then conducts an NHST and, if the test result is "non-significant," runs a few more participants, then conducts another test, then adds a few more participants, and so on until either the test comes up "significant" or the budget for running participants runs out. This all-too-common practice is often justified in terms "efficiency." But it doubles, triples, or even quadruples the probability of Type I error.

For instance, if one runs two groups of 20, tests and fails to attain "significance," then increases the groups by 5 each, then tests again, and so forth, until attaining significance, the Type I error rate is nearly tripled to 13.3%. If one re-tests after each participant added beyond the first 10, the Type I error rate more than quadruples to 22.1%. Worse still, I have heard tell of laboratories in which significance tests were automatically calculated by computer after each new participant's scores were added to the database. The moment bare significance was reached, the computer signaled the researchers and data collection was terminated. The Type I error rate produced by this procedure is astronomical.

This practice goes by names such as "optional stopping" and "sequential testing," but it is by no means the only way in which researchers can contribute to the piling up of *p*-values just below $p = .05$. They can also use several measures of the dependent variable at one time (e.g., multiple tests of anxiety), then report only the measure that reached significance first. Such researchers might believe that are simply looking for the "most sensitive" measure, but what they are often actually doing is selecting the one that first generates a Type I error. Second, researchers often delete from the final write-up whole groups that were non-significantly different from the others, but this also inflates Type I error rates. A third technique is to attempt to "squeeze noise out" of the data by partialing out, *post hoc*, a variety of covariates until one generates a significant relation in the remaining variance between the main IV and the DV (see Simonsohn, Nelson, & Simmons, 2014). When used in combination, such practices can raise Type I error rates dramatically. Indeed, Simmons et al. (2011) ran an actual, empirical study in which they combined them to (seemingly) demonstrate that participants *become a year-and-a-half younger* after listening to the Beatles' song "When I'm Sixty-Four" than when listening to the Wiggles' song "Kalimba." It was an absurd result, to be sure, but imagine what the probable reaction would have been if these very same practices were used produce a more plausible outcome. They would likely be accepted as an authentic "discovery."

I don't believe that many researchers who engage in these questionable research practices see themselves as intentionally manipulating *p*-values. Instead, I would guess that they see themselves as drawing on all of the skill and insight they have gained over years of laboratory experience to extract "subtle" effects that, they imagine, less experienced researchers would not be able tease out. Unfortunately, their concern with "subtle" effects is not matched by an equal concern about generating Type I errors. Most journal editors and reviewers don't seem to have that concern either, though there have been some signs of change recently.

Way back in 2005, John Ioannidis (2005) published an article with the provocative title, "Why Most Published Research Findings are False." Using a broadly Bayesian model of how articles get published, Ioannidis showed that, for a wide range of plausible sample sizes, effect sizes, power levels, etc., it was at least as probable as not that any given article in the published literature reports a false positive result. A critical element of this sort of analysis, however, is what we estimate to be the balance, among all the hypotheses that we actually put to empirical test, between those that are, in reality, true and those that are, in reality, false. For instance, of all the hypotheses psychologists test empirically, imagine that one in five is actually true. We cannot know the correct proportion with any certainty. So, it is a useful exercise to try out a few different plausible numbers to see how the outcome varies.

Although this question and its answer are often attributed to Ioannidis today, the statistician Michael Oakes (1986) included a version of it in his textbook, *Statistical Inference*, back in the mid-1980s. To simplify matters, let us assume that journals publish all and only "significant" findings. Many people are inclined to answer the question of what proportion of the published literature reports Type I errors with a quick, "around 5%," because that is the value to which we conventionally set α (the probability of making a Type I error). But that is incorrect: α is not the proportion of *significant* findings that are Type I errors; α is the proportion of *true null hypotheses* tested that result in Type I errors. In short, it uses the wrong denominator for the question of interest: the number of true null hypotheses rather than the number of significant (and therefore published) findings.

It is easier for most people to understand the correct answer to this conundrum when the numbers are framed in terms of frequencies instead of probabilities (Gigerenzer & Hoffrage, 1995). So, consider the situation in which 5000 substantive hypotheses have been tested, but only 1000 of them (1 in 5) are actually true. In addition, assume that α has been set to .05 and that the power of the 1000 tests of true substantive hypotheses is .50 (a little higher than is typical for psychological research (see, e.g., Cohen, 1962; Sedlmeier & Gigerenzer, 1989; Szucs & Ioannidis, 2016).

Table 1 (adapted from Dienes, 2008) may help to clarify matters: What we do *not* want to do, tempting as it may be, is to rely solely on the numbers that lie in the column under the heading "H_0 true": 200 falsely rejected null hypotheses divided by 4000 total true null hypotheses, which gives us only our α of .05. What we are interested in, instead, are the bold numbers in the row labeled "Reject H_0": 200 Type I errors out of a total of 700 significant (i.e., published) findings, which equals .29. Thus, if 1 out of 5 hypotheses that we test is true, α=.05 and we attain power=.50, then 29% of our published findings are Type I errors.

State of the World				
Decision	H_0 True	H_0 False	Total	Probability of Type I error in the published literature
Accept H_0	3800	500	4300	
Reject H_0	**200**	**500**	**700**	**200/700= .29**
Total	4000	1000	5000	
Proportion rejected	200/4000 = .05 (α)	500/1000 = .50 (power)		

Table 1. Frequencies of 5000 hypothetical research projects, 1000 of which tested true substantive hypotheses. Power = .50, α = .05. If journals published all and only "significant" results, then 200 out of 700 articles (29%) would be reporting a Type I error.

If we were better than that at hypothesizing and, say, 1 in 3 of the hypotheses we tested were true, then the Type I error rate in the published literature would drop to 17%. On the other

hand, if we were worse, so that just 1 in 10 of the hypotheses we test were true, then the Type I error rate in the published literature would rise to nearly half: 47%.

This all assumes that α is set to .05. But, as we have seen (Simmons et al., 2011), certain common but questionable research practices shift the true probability of making a Type I error to higher values. Think about the scenario in which many psychologists test two groups of 20 participants, do a test, which fails, then add successive batches of 5 more, testing after each batch until attaining significance. According to Simmons et al., that raises the Type I error rate among the studies conducted to 13.3%. Assuming, as we did initially, that 1 out of 5 hypotheses tested are true and that power = .50, that would bring the Type I error rate in the published literature to 52%: more than half, just as Ioannidis (2005) predicted. If researchers were re-testing the data after every participant beyond the first 10, then the published Type I error rate would rise to a terror-inducing 64%: nearly two-thirds of published articles would be reporting Type I errors!

There was, of course, an earlier version of this "crisis" around significance testing. It was in the 1960s – more than a career ago. People like Jum Nunnally (1960), William Rozeboom (1960), Jacob Cohen (1962), David Bakan (1966), Paul Meehl (1967, 1978), David Lyyken (1968), and John Tukey (1969) warned us of many of the issues that haunt us today. Nunnally told psychologists that "in the real world, the null hypothesis is almost never true, and it is usually non-sensical to perform an experiment with the *sole* aim of rejecting the null hypothesis…. [I]f the null hypothesis is not rejected, it is usually because the *N* is too small " (Nunnally, 1960, p. 643). Instead, according to Nunnally, we should be estimating effect sizes, though he did not use exactly that modern phrase: instead he wrote "how much of the total variance is explained by particular classifications?" [i.e., differences among groups] (Nunnally, 1960, p. 647). He then went on to discuss the use of confidence intervals, not around means but, rather, around the effect size estimates.

Speaking out even more strongly, Rozeboom regarded the entire null hypothesis-testing testing procedure to be a "fallacy" and a "quorum of embarrassments" (Rozeboom, 1960, p. 417). NHST is underpinned by what he regarded as a plainly false assumption that the aim of an experiment is to make a stark decision to either reject or accept the null hypothesis when, instead, one should be merely adjusting one's degree of belief in the probability of a hypothesis, not deciding for or against it. It is, perhaps, surprising to see a psychologist advocating for the Bayesian statistical view some 60 years ago, but it is worth recalling that David Bakan (1953) had featured "inverse probability" in a major psychological periodical several years earlier. The article that is usually regarded as the primary initial presentation of the Bayesian view to psychologists (Edwards et al., 1963) would appear just three years after Rozeboom's article. In addition, Sharon Bertsch McGrayne (2011, pp. 163–175) reported in her history of Bayesian theory, that John Tukey regularly operated in a Baeysian mode, but that he did so in research that the U.S. Gov't regarded as "classified," so very few others were aware of his activities in this vein.

Returning to Bakan, in addition to his promotion of Bayesian analysis, he provided a trenchant critique of null hypothesis testing. In the mid-1960s, he reminded psychologists that "the test of significance does not provide the information concerning psychological phenomena characteristically attributed to it; and… furthermore, a great deal of mischief has been associated with its use" (Bakan, 1966, p. 423, reprinted in 1967, chapter 1). He repeated Nunnally's assertion that the null hypothesis is nearly always false, bolstering it with a story in which he had once analyzed multiple test scores collected from 60,000

people: "Every test came out significant. Dividing the cards by such arbitrary criteria as east vs. west of the Mississippi River, Maine vs. the rest of the country, North vs. South, etc., all produced significant differences in means" (Bakan, 1966, p. 425). Bakan also noted that the widespread policy of journals to reject articles in which significance had not been obtained was probably leading to a glut of uncorrected Type I errors in the scientific literature, noting the then-recent (now classic) study of statistical power by Jacob Cohen (1962) which showed that the average probability of detecting even a medium sized effect was less than 50% in *Journal of Abnormal and Social Psychology*. It is worth noting that this was all more than a decade before Robert Rosenthal published his famous "file drawer problem" article (Rosenthal, 1979).

David Lykken (1968) took this same set of concerns in a somewhat new direction, describing how issues related to random sampling, especially in difficult-to-obtain clinical populations, could have a profound effect on the results obtained. In the final analysis, he concluded that

> the finding of statistical significance is perhaps the least important attribute of a good experiment; it is *never* a sufficient condition for concluding that a theory has been corroborated, that a useful empirical fact has been established with reasonable confidence – or that an experimental report ought to be published. (Lykken, 1968, p. 159).

Back in the 1960s, most psychologists simply ignored these multiple caveats, if they were aware of them at all, and continued on much as they had before. For those whose main occupation was to develop of coherent account of scientific methods, though, the situation in psychology seemed dire. The renowned philosopher of science, Imre Lakatoš, after reading Lykken's article and another by Paul Meehl (discussed below) drew this dark conclusion:

> One wonders whether the function of statistical techniques in the social sciences is not primarily to provide a machinery for producing phoney [*sic*] corroborations and thereby a semblance of 'scientific progress' where, in fact, there is nothing but an increase in pseudo-intellectual garbage. (Lakatoš, 1970, p. 176n)

But this was all too abstract for most psychologists back then, and these warnings were largely ignored.

Why did so many psychologists treat these bleak appraisals so casually? It is often assumed that it was largely a matter of poor statistical education: psychologists-in-training were generally provided with a fairly narrow range of statistical tools, and relatively little in the way general probability theory with which they might come to the same conclusions as the critics had. In addition, most psychologists (journal editors included) were interested in pursuing the problems of mind and behavior, not in delving into what seemed like arcane statistical issues. Statistical analysis in psychological research seemed to be "working," more or less, so why go to all that trouble merely to upset what, to all appearances, was a relatively stable cart?

Most of the improvements in statistical training for psychologists over the second half of the 20th century came in the form of learning more complex methods, enabling researchers to handle more variables and more complex relations among them, not in the form of a

stronger grounding in basic statistical principles. I believe, though, that the problem goes deeper than a simple failure of education and, consequently, that the problem cannot be solved merely by more intensive statistical instruction. It is not simply that psychologists are undertrained in statistics. It is, rather, that the statistical procedures we use, limited as they might be, are actually well-suited to the weak theories that psychologists typically produce. And, by means of this feedback loop, the range of theories that we typically produce is narrowed by the statistical procedures that we are expected to use. This insight was outlined, in essence, more than 50 years ago by Paul Meehl (1967) in an article was mostly ignored by psychologists, likely because it was published in the journal *Philosophy of Science*.

Meehl noted that, in the "hard" sciences, researchers do not merely reject the null hypothesis in order to declare a success for their favored substantive theory. Instead, their theories are expected to make specific point predictions, and the data collected must strongly confirm that specific prediction, not just fail to support a "straw man" prediction of zero. What is more, Meehl continued, the requirement that data confirm point predictions means that, as measurement procedures gradually become more precise and the error interval around it narrows, a prediction that had once been (weakly) confirmed by a highly variable measurement procedure might later become disconfirmed as the range of measurement error lessens. This, Meehl argued, is what forces science to improve and progress. For a visual representation of this, see the two brackets below the line on the right-hand side of Figure 3: In the upper one, a less precise measurement captures (barely) the value predicted by H_1. The lower one, however, shows that a more precise measurement with exactly the same mean value fails to capture the value that H_1 predicts. Increased precision makes accurate prediction more difficult.

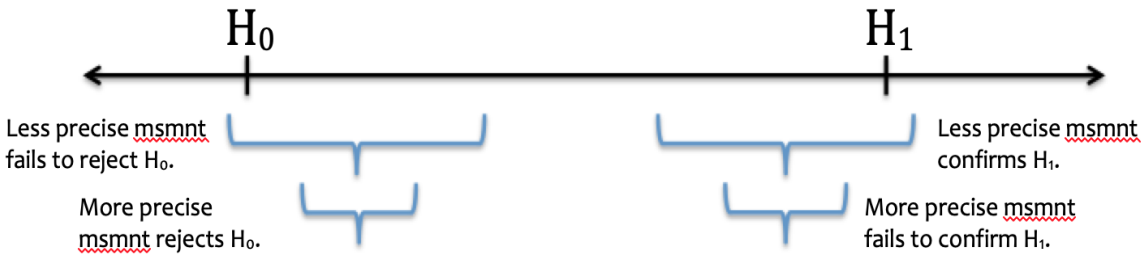

Figure 3. The right-hand brackets show H_1 becoming more difficult to confirm as the measurement procedure becomes more precise. The left-hand brackets show H_0 becoming easier to reject as the measurement procedure becomes more precise.

By contrast, when one relies on null hypothesis testing, the improvement of measurement has precisely the opposite effect: as improved measurement procedures lead to a narrowing of measurement error, a null hypotheses that failed to be rejected earlier might now come to lie in the rejection region, without the mean of the data moving at all. See the brackets on the left-hand side of Figure 3: The upper one shows a less precise measurement failing to reject the H_0. The lower one shows a more precise measurement with the same mean rejecting H_0.

Meehl regarded this as a looming disaster for psychology and, about a decade later, he produced a simplified version of the argument for consumption by psychologists

(unfortunately buried among a number of other methodological criticisms). Pulling no punches, Meehl wrote to his psychological colleagues:

> I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology. (Meehl, 1978, p. 817)

Alas, the great bulk of the discipline ignored this call to arms. John Tukey had summed up the matter in a brief but vivid simile in the late 1960s: (paraphrasing slightly), You're never going to get a science like Newtonian physics out of conclusions like, "If you pull on it, it gets longer" (Tukey, 1969, p. 86). But that is exactly the kind of conclusion that significance testing provides: it predicts the direction of the effect (viz., longer) without predicting its size. Predicting the size of the effect – i.e., the exact location of the outcome – is a much stronger form of prediction because it rules outs many more possible outcomes than merely predicting its direction does. Unfortunately, Tukey's illustration of the principle never caught the attention of many psychologists either, though it was later quoted approvingly by Jacob Cohen (1994).

To me, it now appears obvious that it was all the way back in 1960s that we predisposed ourselves to the replication "crisis" we are experiencing today. Null hypothesis testing was *always* an extremely weak criterion for scientific truth, one that we seized upon over-eagerly because it seemed at the time to give psychologists the scientific *bona fides* they so desperately craved. What is more important, psychologists continued to rely on it because, first, it became the foundation on which they built their academic careers. Rejected nulls increasingly resulted in journal articles. Journal articles increasingly resulted in graduation, employment, and tenure. Tenure resulted in careers. Those who did not follow this path were selected out of the discipline in a virtually Darwinian fashion and, thus, the "Null-Rejectors" became the dominant species of the ecosystem known as "psychology."

Second, most psychologists have no idea of how to construct a theory that would make point predictions. To be clear, I don't mean just that most have never been taught the skills of scientific theory construction – though that is true as well. What I mean is that psychologists have never developed the basic conceptual foundation – if it is indeed possible to couch psychology in terms of such a foundation – that would enable them to make such predictions. Imagine that you have developed some new "energy drink" that you believe will improve people's attentional focus. By how much? For how long? The issue here is not solved by simply taking empirical measurements of the average improvement in attention and the duration of that increase. It is to create a theory that includes a concept of attention that can be manipulated in some sort of formal, rigorous way – e.g., mathematical, logical, grammatical, game theoretic, computational. The problem is, essentially, that we still don't have the psychological equivalents of mass and velocity and force, and a fundamental formula that tells you how they related to each other, like Newton's $F=ma$.

To conclude, it is certainly a problem that significance testing is too weak a criterion of scientific testing truth AND, over and above this, that significance testing is too easily distorted by researchers who usually mean no harm, but who mostly do not fully understand its limitations, and who are subject enormous extraneous pressures to be "productive" (crudely defined as publishing articles and collecting citations). But this is not the

fundamental problem. The fundamental problem is that we have never developed the precise theoretical apparatus that would allow us to *escape* from significance testing, so we continue to lean on it as a kind of scientific crutch – a sign to the rest of the world that we are "scientific." Ironically, though, many critics, both inside and outside of the discipline, now seem to have noticed that the crutch is broken and fails to support our truth claims.

Worse still, it is not like we can simply acknowledge the problem and start developing the new kinds of theories that we need instead. In the vast majority of psychological domains, we don't know where to begin in creating theories that could make point predictions about attention or depression or attachment or conscientiousness or intelligence, etc. even possible. Until we figure out how to solve *that* problem – or collectively decide that it is insoluble – I think that we are likely to remain trapped inside the significance testing box.

References

Bakan, D. (1953). Learning and the principle of inverse probability. *Psychological Review, 60*, 360–370. https://doi.org/10.1037/h0055248

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423–437. https://doi.org/10.1037/h0020412

Bakan, D. (1967). *On method*. Jossey-Bass.

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. https://doi.org/10.1037/a0021524

Bem, D. J., Tressoldi, P. E., Rabeyron, T., & Duggan, M. (2015). Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of radom future events. *F1000 Research, 4*, 1188. https://doi.org/10.12688/f1000research.7177.2

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145–153. https://doi.org/10.1037/h0045186

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cumming, G. (2009). *Dance of the p-values*. https://www.youtube.com/watch?v=ez4DgdurRPg

de Winter, J. C. F., & Dodou, D. (2015). A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ, 3:e733*. https://doi.org/10.7717/peerj.733

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Palgrave.

Ebersole, C. R., & 63 others. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242. https://doi.org/10.1037/h0044139

Engber, D. (2017, June 7). *Daryl Bem Proved ESP Is Real. Which Means Science Is Broken.* Slate Magazine. https://slate.com/health-and-science/2017/06/daryl-bem-proved-esp-is-real-showed-science-is-broken.html

Etz, A. (2015, August 30). The Bayesian reproducibility project. *The Etz-Files*. https://alexanderetz.com/2015/08/30/the-bayesian-reproducibility-project/

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124https://doi.org/10.1371/journal.pmed.0020124

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Klein, R. A., & 50 others. (2014). Investigating variation in replicability: A "'Many Labs"' replication project. *Social Psychology, 45,* 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., & 173 others. (2018). Many labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490. https://doi.org/10.1177/2515245918810225

Lakatoš, I. (1970). Falsification and the methodology of science research programmes. In I. Lakatoš & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge University Press.

Lakens, D. (2014, September 30). What p-hacking really looks like: A comment on Masicampo & LaLande (2012). *The 20% Statistician*. http://daniellakens.blogspot.com/2014/09/what-p-hacking-really-looks-like.html

Leek, J. T. (2017, October 2). Creating an expository graph for a talk. *Simply Statistics*. https://simplystatistics.org/2017/10/02/creating-an-expository-graph-for-a-talk/?utm_content=buffer76db6&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159. https://doi.org/10.1037/h0026141

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal of Experimental Psychology*, *65*, 2271–2279. https://doi.org/10.1080/17470218.2012.711335

McGrayne, S. B. (2011). *The theory that would not die: How Bayes' Rule cracked the Enigma Code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*. Yale University Press.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103–115. https://doi.org/10.1086/288135

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Nunnally, J. C. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, *20*, 641–650. https://doi.org/10.1177/001316446002000401

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Wiley.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, *57*, 416–428. https://doi.org/10.1037/h0042040

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316. https://doi.org/10.1037/0033-2909.105.2.309

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p Values for Testing Precise Null Hypotheses. *American Statistician*, *55*(1), 62–71. https://doi.org/10.1198/000313001300339950

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*, 1359–1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547. https://doi.org/10.1037/a0033242

Szucs, D., & Ioannidis, J. P. A. (2016). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *BioRxiv Preprint*. https://doi.org/10.1101/071530

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*, 83–91. https://doi.org/10.1037/h0027108

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432. https://doi.org/10.1037/a0022790

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

About the author

Christopher Green is a professor in the Historical, Theoretical, & Critical Studies of Psychology graduate program at York University in Toronto, Canada. He hold PhDs in cognitive psychology and in the philosophy of science, both from the University of Toronto. Much of his research has been on the history of psychology, including the 2019 book, *Psychology and Its Cities* (Routledge). He has pioneered digital methods of historical research, co-authoring more than a dozen articles that explore the development of the intellectual structure of psychology, using large databases of the journal literature through time. He is now bringing digital methods to bear on the statistical and methodological origins and of the "replication crisis" in psychology. He has also written on the (lack of) theoretical unification in psychology, the aesthetics of the golden section, and the history of baseball. He has served as president of the Society for the History of Psychology, editor of the journal of the *History of the Behavioral Sciences*, and is the incoming editor of *History of Psychology*.

Contact: Department of Psychology, York University, Email: christo@yorku.ca

ORCID: https://**orcid**.org/0000-0002-6027-6709