Why is the statistical revolution not progressing?

Vicious cycle of the scientific reform

THIS IS A PREPRINT

The manuscript has not yet been peer-reviewed.

Michał Białek¹, Michał Misiak^{1,2}, and Martyna Dziekan³

1. University of Wrocław, Institute of Psychology
2. University of Oxford, School of Anthropology & Museum Ethnography
3. Adam Mickiewicz University, Faculty of Psychology and Cognitive Science

Author Note

The data used in this project are publicly available (<https://osf.io/ubyj2/>).

We have no conflict of interest to disclose.

Acknowledgments: We would like to thank Daniël Lakens for his feedback on the project and comments on its first draft. Michał Misiak was supported by the scholarship of the Foundation for Polish Science (START) and by the scholarship of the National Science Centre (2020/36/T/HS6/00256).

Corresponding Author: Michał Białek, Institute of Psychology, University of Wrocław, ul. Dawida 1, 50-527 Wrocław, Poland; michal.bialek3@uw.edu.pl

Abstract

To analyze the results of the research, behavioral scientists widely use a statistical rule that sets the significance level to 0.05. Recently, two recommendations on how to improve statistical inference were published: to redefine statistical significance to 0.005, and to select and justify the alpha. We analyzed the empirical work that cited the original recommendation papers, as well as the papers published by the scientist that co-authored the publications. About half of the numerous papers citing these recommendations adhered to them already in the first year since their publication. What is striking, the original authors that proposed the recommendations followed their own recommendations only in 6% of their empirical work. We surveyed the authors asking them to identify major obstacles they experienced while trying to implement their own recommendations, and obstacles they think others could expect or experience.

The Problem

To many behavioral scientists, statistics is like driving a car: they do not have to know how the engine works to participate in the traffic, you only need to learn the traffic rules, and how to operate the car. In science, a similar logic applies to conducting statistical tests. Scientists learn the statistical traffic rules—like those suggesting when a test should be used or what is the inference strategy given a test result. Then, one can simply click some icons in dedicated software, get some output, and search for the magical number, the P value. From time to time, cars undergo a check in a garage to assure it's safe to drive. Recently, however, “statistical mechanics” raised their voice to suggest that the procedure which scientists use to make inferences is broken, and it is not safe to science without fixing it.

To understand what went wrong, we need to consider one of the most popular rules behavioral scientists use. When deciding whether the data support the hypothesis, there are two potential errors they might commit: supporting a false hypothesis (Type I error) and failing to support the true hypothesis (Type II error). The probability of committing the Type I error is captured by the alpha level of a test: the probability of observing the same or more extreme results given there is no true effect. One never knows if a decision to accept or reject a hypothesis will be correct in a particular case, but if the P value is lower than the alpha level, we will be correct most of the time in the long run¹. This is the best we can get from observing the world and analyzing noisy data from it. This logic is called Null Hypothesis Significance Testing (often acronymized to NHST). For behavioral science, the accepted level of surprise (α) is assumed to be 5%. To illustrate the NHST: if two groups differ in their test score (i.e., one $\bar{X}_1 = 75$ and $\bar{X}_2 = 90$, therefore $\Delta = 15$), and the t -test returns is, say, $P = .02$, researchers would reject the null hypothesis thinking that there is only 2% chance to observe the difference at least as large (i.e., $\bar{X}_2 - \bar{X}_1 \geq 15$) had the groups be truly performing

identically. Rejecting the null allows scientists to accept an alternative hypothesis that the groups differ significantly.

Neyman-Pearson approach to scientific inference used a significance threshold to control for cumulative accuracy of scientific claims, i.e., that if all claims would be used with $P < .05$ criterion, scientists would be wrong not more than in 5% cases. Behavioral scientists know the $P < .05$ rule and follow it in most of their empirical work. However, some scientists claim the 5% threshold used to infer significance is too liberal, resulting in too many false discoveries. Indeed, some branches of science decided on far more conservative α . For example, in physics, the accepted Type 1 error rate is set to $5.7 * 10^{-7}$. The next issue is that the evidential value of significant results under a 5% threshold is rather low, corresponding roughly to evidence of only 3:1 in favor of the tested hypothesis over the null ²—and this only while assuming all tests were conducted honestly and correctly, which as in all human activities, is obviously not true. Furthermore, because scientific journals usually prefer to publish new and surprising discoveries, scientists are incentivized to run studies that show significant test results. The method of significance testing can be bent, to obtain $P < .05$. Doing so increases the false positive rate. The widespread statistical practices to “help” the P value to cross the level of significance are now known as p-hacking^{3,4}. For example, researchers could tackle a dataset with multiple statistical tools, until one analysis returns a P value below the level of 0.05. They could also operationalize the variables with multiple methods, and only report the ones that “worked”. Having a non-significant test result, researchers could also add participants until repeated analysis on the new dataset will be statistically significant. In most of the cases, these practices are not intended to be fraud, but could be learned by observing successful peers and their conduct; a secret knowledge of how to make things work (see now infamous “tutorials” by Bem⁵ or Wansink⁶). The rule of using a significance level of 0.05 in behavioral science has been questioned and blamed for the flood

of false-positive findings^{7,8}. To resolve this problem, two large groups of scientists recently proposed their solutions.

The Solutions

Two large teams, one lead by Daniel Benjamin² and the other lead by Daniël Lakens⁹, discussed ways of improving the rules associated with inferring statistical significance. The first group postulated lowering the threshold of significance from 0.05 to 0.005. The second group argued that replacing one threshold with another is not a way to go, and instead, researchers are free to select any alpha level, as long as they can justify it.

Let us have a closer look at the first suggestion. Moving α from 0.05 to 0.005 increases the evidence strength from 3:1 to at least 14:1 in favor of the hypothesis against the null. Another advantage of the proposal is its ease of implementation: it simply updates a rule everyone knows. Yet, there are costs associated with the new rule: maintaining the sample size as previously would increase the type 2 error, i.e., it is more likely to fail to observe an effect that truly exists; decreasing the alpha level but retaining the same type 2 error would require 70% larger sample sizes.

The second suggestion claims scientists should a priori select an α threshold that will allow them to decide whether the hypothesis is supported. Anything goes, given the justification of the alpha level is sound. You are doing exploratory research, or study a small population? Try $\alpha = .1$. You are doing confirmatory research, or maybe you have a large dataset? Try $\alpha = .01$. This flexibility in setting inference strategy is appealing but does not follow the simple and rigid rules we believe scientists seem to prefer to apply in their research. The recommendation burdens the editors and reviewers with additional requirements to check the justification of the α , without them necessarily possessing the required expertise.

What is critical in both of the two recommendations is that the decision to adopt a particular α -level should be made a priori, that is before looking at the data, and preferably at

the stage of designing an experiment. Only this serves the purpose of controlling the long-run error rates. Modifying the α after the results are known does not provide this benefit – one could report the results using a stricter threshold for significance only after knowing the results pass it, merely creating the illusion of tight control of error rates. Another obstacle is that implementing both recommendations is only feasible in a preregistered project because the α and corresponding β should be determined before the data are collected. And, although gaining popularity, preregistrations are not yet a default¹⁰.

Both recommendations received huge interest. Their Altmetric scores as of December 2020 being: 878 for the paper that postulated lowering the alpha level and 167 for the paper that postulated justifying the alpha level. A query in Google Scholar at the same time returns a citations count of 1315 and 234, respectively. As intuitively expected, the simpler rule to move α to 0.005 that can be unreflectively followed was far more discussed and cited. But even the less popular paper gained so much interest, that if a journal published only papers like this one, it would be among the world's most-cited journals. To wrap up, it seems that the new rules of statistical inference attracted much interest. This is a necessary condition for a lasting change, but sole citations are not a marker sufficient enough. In the end, scientists have to implement the new rules in practice. But do they?

The Implementation

The popularity of both suggestions on how to change statistical rituals promised a rapid implementation in research practice. However, we were concerned that the correspondence between citation and implementation can be poor, once we considered a case of The False-Positive Psychology paper⁷. In it, Simmons, Nelson, and Simonsohn – among other recommendations – urged scientists to consider statistical power when designing experiments. They argued that researchers should include a minimum of 50 participants per cell when investigating large effect sizes, and even more participants when expecting to

observe smaller effects. Despite its popularity, the paper was often miscited, e.g., to support claims that one should not collect more than 50 participants per condition¹¹. As Simmons and colleagues acknowledge: although a citation necessarily indicates that a researcher was influenced to reference the article, it does not necessarily indicate that the researcher was influenced to carry out the intentions of the article¹¹. We wondered whether the same applies to the two papers recommending the change in the P value rule.

We verified whether any of the postulated recommendations lead to change in scientific practice. To this end, we screened $N = 3217$ papers written in English with two goals: (1) to see whether the authors who cited at least one of the two discussed publications followed the recommendations and (2) to see whether the authors of the discussed recommendations adhered to their own suggestions. We initially intended to predict the implementation of the recommendations using information about the authors (their academic position, citation count, H-index) or the work (whether it was preregistered, number of co-authors). Yet, the skewed distribution of dependent variables and generally low number of papers/scientists that implemented the new rules made such analyses unreliable. Raw data are available online (<https://osf.io/ubvj2/>).

Before we proceed to discuss the results of the survey on citations, we need to clarify few things. That is, we do not intend to argue the high or low level of implementation just after the recommendations are published are indicative of whether the recommendations are sound, or will be successfully implemented in the future. This part of our manuscript is rather a starting point for more direct inquiry and discussion about the prospects of revolutionizing the statistical practice in behavioral sciences, and obstacles preventing it from happening. For our query, we only searched through the Google Scholar database. Still, it is the most comprehensive database, includes preprints, and is suitable for meta-analyses¹². Our data cover 2017-2020. Many of the projects were published as journal articles or as preprints

reporting experiments designed and conducted years before. Those started after the publications of the recommendations have probably not been published yet. So, our query does not necessarily cover the extent to which the rules are adopted in ongoing or future research, nor does it suggest that the authors are unwilling to follow their own recommendations. Instead, we wanted to learn about the practice of implementing these recommendations. Specifically, we wanted to see how widespread these rules are in already completed projects, and the experience the authors could have collected while adopting them to their ongoing projects.

Let us have a look at the effect of the publication that postulated to lower the alpha level first. At the time we started our query, Google Scholar identified 737 papers written in English that cited this paper, with only 294 papers reporting empirical work that used NHST (that is non-Bayesian) approach. A total of 169 (57%) empirical papers lowered the alpha threshold to 0.005 to decide on statistical significance. This is a large index of implementation. However, some of these citations are “enforced citations”, where the authors explicitly admit they were asked by the editors or reviewers to change their α to 0.005. So, the editors and reviewers are sometimes “helping” authors to adopt the new recommendations *after* the study was conducted. This seems to contradict the core reason for adopting the alpha level, which is a more severe test of a hypothesis given a priori decided criterion of inference with a sufficiently powered experiment. Changing inference criterion once the results are known does not seem to be serving this purpose well. One additional observation is that the authors who did not lower the threshold, often cite this paper when observing $P < .005$ and claiming it would be also significant with this new, stricter threshold. So, they used the “lower the alpha” recommendation to rhetorically strengthen support for their conclusion. Thus, some of the applications of the criterion do not serve the purpose of the recommendation, but still, populate the literature with references to it. In principle, this is a good thing, because other

researchers become familiar with the new approaches to statistical inference. However, we are worried that such misrepresentation of the criterion (used *post hoc* rather than *a priori*) may effectively distort the main reason the recommendation was suggested.

Next, we looked at the authors who proposed lowering the α -level to evaluate the extent they implemented their own recommendation. The 72 authors of the recommendation published 1280 unique papers between the publication date and June 2020, averaging to 18 papers per author. Among these papers, only 365 reported empirical work, with only 37 papers (10%) actually implementing the suggested rule of $P < .005$. Of the few that did, 14 papers came from fields in which stricter thresholds are already a standard practice (e.g., in genetics). Similar rates of adopting the $P < .005$ rule can be found when limiting our query to papers in which the original authors are first, corresponding, or last authors of new work (23 out of 174 papers, or 13%).

Now, for the second suggestion to improve statistical inference—justify the alpha level you are using. As we said, it requires researchers to reflect on the rules instead of simply blindly obeying them. This paper was far less popular than the “Redefine” one, receiving less media attention and fewer citations. Its implementation was even less widespread. We inspected 121 papers citing the work, out of which only 19 were empirical and used the NHST approach, and 8 (or, 42%) of these papers indeed justified their α -level. Out of the 1079 unique papers of the 88 original authors of this recommendation (12 papers per author), 559 were empirical with the NHST approach with only 18 (3%) of them justifying the selected α -level.

Taken together, out of 1237 empirical papers citing any of the two recommendations we inspected that used the NHST approach, 232 (19%) implemented one of the recommendations in praxis. This can be seen as a large number, especially considering the short period between the publication of the recommendations and the time of our query; yet

this is just a drop in a sea of all empirical work published during that time. The share in the published empirical work grew from 2019 to 2020 (Fig. 1). Hence, it is plausible to expect the number of published work that implements any of the recommendations to grow.

Fig. 1. The proportion of papers published by the authors of the recommendation (“authors”) and of papers that cited the recommendations (“citations”) that implemented the recommendations in practice.

Note. The data for 2020 only cover the months from January to June.

A large number of citations that did not follow the recommendations suggests that researchers are aware of the ongoing debate about the α but decided to follow the traditional way for now. This implies that once the default would change, they could be open to adopting

the recommendations. It is also possible that there are objective obstacles that prevent researchers from adopting the recommendations. To gain some insight into this issue, we decided to survey the authors of these recommendations to learn about their own experience with such publications and about what they think stops others from adopting these recommendations.

The Obstacles

We invited all the authors of the recommendations to participate in a short survey ($N = 160$), asking them two open-ended questions: (1) what is the reason you implemented/decided not to implement this recommendation in your work?; (2) what is your opinion, what are the obstacles you believe are preventing other researchers from implementing this recommendation? We received quite some responses ($n = 33$ for “Redefine” and $n = 51$ for “Justify”) and decided to analyze them qualitatively, by broadly categorizing them into groups of arguments. The full list of anonymized responses can be found in the OSF folder (<https://osf.io/ubvj2/>).

Why did the authors not follow their recommendations?

The most common response from the authors of the “Redefine” suggestion ($n = 11$) to why they decided not to implement the recommendation was that their coauthors decided against it. Not much more is known about the specific reasons in particular cases, but we think that the status quo in the field plays a significant role here. Psychological research suggests people expect good reasons to depart from the status quo but do not require similarly strong arguments not to honor it¹³. That is, if the arguments for and against a particular solution are of similar strength, most of the people will rather stick to the status quo¹⁴. Here, the arguments for adopting the recommendation have to be really strong to convince a team to risk a new statistical approach in their project. Once the P value < 0.005 would be more popular, or better - the default, more researchers would be supportive of this solution. That is,

we do not believe majority of researchers have strong convictions about why 0.05 is the best alpha level, better than other potential levels, other than “this is the default”.

The second most common explanation was that the interpretation of the results ($n = 5$) was independent of the α , thus, there was no need to bring this issue in the paper. This is potentially problematic, given that the goal of changing the α is to do this a priori, preferably as a part of preregistration. Only then the adopted threshold is truly impacting the conclusion about the hypothesis. Otherwise, researchers risk a bias which—by analogy to publication bias—we labelled “the alpha-bias”: if $P < .005$ researchers will admit their results are significant at the new stricter level, but if $.005 < P < .05$ they may not discuss this issue and proceed with standard statistical inference. All other reasons provided by researchers for why they did not implement their recommendations in the surveyed projects were hard to cluster and were rather case-specific, e.g. that a project aimed to show null-effect, and stricter α would harm the goal rather serve it.

Some of the authors replied to the invitation via email rather than taking the survey. The two most common arguments for low implementation mentioned in the email conversations were: the NHST is not the way to go in general, and therefore the authors prefer alternative inference approaches (e.g., Bayesian), or that their recently published work reports data collected much before the publication of the “Redefine” paper.

A similar picture to the one described above is painted by researchers who responded to the survey on the “Justify” recommendation. Again, the most common reason for not adopting the recommendation was that it was hard to convince the coauthors ($n = 18$). Furthermore, $n = 17$ responses referred to the timing: some of the papers reviewed by us reported projects that started before the recommendation was published. Some of the authors ($n = 9$) reported pushback from reviewers or editors. Finally, $n = 7$ responses referred to

problems with the justification of the α , e.g., not having enough knowledge to ground the justification.

What both teams had in common is that they received pushback from their collaborators as well as from the reviewers and editors. It seems that scientific society is not yet ready to adopt the recommendations. The majority of researchers are unfamiliar with the recommendations, and from those who know about the suggested improvements, many refuse to follow them. What are the reasons that prevent other scientists from adopting the recommendations? We asked the authors of the recommendations to predict what they could be, or what they know these are.

What prevents other scientists from adopting the recommendations?

Let us again start with the “Redefine” recommendation. The authors speculate that what holds others back from attempting to apply this recommendation are: it will be harmful to their careers ($n = 9$), such research requires larger samples and funds ($n = 8$) and that this is only feasible given a wide consensus on this rule ($n = 8$). Further responses referred to decreased chances of finding an interesting result ($n = 5$), cycling back to the power issue. These responses suggest that the authors are fully aware of how hard it will be to convince other researchers to adopt the recommendations, even when they may privately agree on them. More money required to run a study translates into publishing fewer papers, which in turn leads to fewer chances to secure a stable job in academia and to successfully apply for additional funds from granting agencies. Although improving statistical inference could benefit society in the long run, it seems that implementing new standards is perceived as harmful for individual scientists.

For the “Justify” team, $n = 20$ responses referred to the lack of materials helping scientists to justify their choices. That is, they think scientists need a tutorial in how to justify their choices, which argument is sound and which is not. This is the main argument raised by

the team. Others are similar to the ones of the “Redefine” team: reviewers and editors are not yet ready to accept the deviations from the default 0.05 criterion ($n = 6$), and that researchers may be worried the recommendation may hinder the publication process ($n = 5$). Related to these points, $n = 4$ responses mentioned the strength of conventions that prevent scientists from adopting new rules and the lack of interest in implementing these rules by journals and funding agencies. Next, $n = 4$ responses referred to strict word counts not allowing for discussions of the data analysis, thus limiting the ability of researchers to justify their selected alpha. Last but not least, $n = 4$ responses mentioned difficulties in convincing co-authors to justify their alpha rather than going with the default.

In general, the respondents admit to experiencing a pushback. They struggle to negotiate to implement new standards with their co-authors and expect to meet with resistance within journals in the field. That is, people do not adopt the new recommendations, because it is not an approved standard in their area; it is not an approved standard in their area because no one is changing their old habits. Hence, the situation looks like a book example of a vicious cycle.

Conclusions

The recurring statistical revolutions reminded us of hockey helmets. Until 1979, hockey players would not use any head protection, resulting in many dangerous injuries. Yet, players were willing to risk their health, because wearing a helmet would limit their ability to see, hear and would increase overheating, altogether negatively affecting chances for their team to win a game. Thus, although players were aware that helmets are what they all need, they refused to wear them because no one else was wearing them. Once helmets became mandatory, all players were protected and negatively affected to the same extent, benefiting their health, and not affecting their relative chances for a victory.

The parallel between statistical revolution and hockey helmets is striking. To meet the new, stricter threshold for significance, one requires much larger sample sizes (if keeping the same power) or suffers reduced power (if keeping the same sample sizes). This then translates into a lower number of studies one can successfully run, and fewer papers one can publish. As a result, this affects their careers, where publish-or-perish is a widely accepted rule. So, as with hockey, many scientists agree that some reforms must be made (many citations to the two recommendations), but they usually refrain from implementing these recommendations fearing to be the first to suffer its negative consequences (very few papers that adopt the stricter significance level or justify the alpha).

Our recommendations, based on the feedback from most reform-invested scientists who co-wrote the two papers postulating a change in statistical procedures is that journals and funding agencies should request P values to be adjusted or justified. Only that way the statistical reforms will become known, and perceived as state-of-the-art, which in the long run will transform into the default. If it worked with effect sizes, required to be reported along with the P value^{15,16}, we hope it can work with other standards that increase the quality of our statistical inferences.

References

1. Lakens, D. The Practical Alternative to the p Value Is the Correctly Used p Value. *Perspect. Psychol. Sci.* **16**, 639–648 (2021).
2. Benjamin, D. J. *et al.* Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6 (2018).
3. Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. & Jennions, M. D. The extent and consequences of p-hacking in science. *PLoS Biol.* **13**, e1002106 (2015).
4. Brodeur, A., Cook, N. & Heyes, A. Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics. *Am. Econ. Rev.* **110**, 3634–3660 (2020).
5. Nadon, R. QUEST Blog: Planes that Don't Land, ESP and Bad Science. <https://www.bihealth.org/en/notices/quest-blog-planes-that-dont-land-esp-and-bad-science> (2018).
6. Wansink, B. The Grad Student Who Never Said 'No'. *HEALTHIER & HAPPIER* <https://web.archive.org/web/20161215112231/http://www.brianwansink.com/phd-types-only> (2016).
7. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
8. Ioannidis, J. P. Why Most Published Research Findings Are False. *PLoS Med.* **2**, e124 (2005).
9. Lakens, D. *et al.* Justify your alpha. *Nat. Hum. Behav.* **2**, 168–171 (2018).
10. Nosek, B. A. & Lindsay, D. S. Preregistration becoming the norm in psychological science. *APS Observer* (2018).
11. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-Positive Citations. *Perspect. Psychol. Sci.* **13**, 255–259 (2018).

12. Gehanno, J.-F., Rollin, L. & Darmoni, S. Is the coverage of google scholar enough to be used alone for systematic reviews. *BMC Med. Inform. Decis. Mak.* **13**, 7 (2013).
13. Anderson, C. J. The psychology of doing nothing: Forms of decision avoidance result from reason and emotion. *Psychol. Bull.* **129**, 139–167 (2003).
14. Gerasimou, G. Asymmetric dominance, deferral, and status quo bias in a behavioral model of choice. *Theory Decis.* **80**, 295–312 (2016).
15. Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front. Psychol.* **4**, (2013).
16. Sullivan, G. M. & Feinn, R. Using Effect Size—or Why the *P* Value Is Not Enough. *J. Grad. Med. Educ.* **4**, 279–282 (2012).