Journal of Expertise
2019. Vol. 2(4)

ISSN 2573-2773

Retiring Statistical Significance from Psychology and Expertise Research

Guillermo Campitelli
Discipline of Psychology, Murdoch University, Australia

Correspondence: Guillermo Campitelli: Guillermo.Campitelli@murdoch.edu.au

Abstract

Eight hundred and fifty-four researchers were signatories of the article "Retire statistical significance" published in *Nature* in March 2019. I was one of them, and in this article, I expand on why retiring statistical significance would improve research in general, and psychology and expertise research in particular. The proposal involves eliminating the binary statistical significance decision in a single study and adopting the collective effort of accumulating evidence over several studies. When statistical significance is removed, the file drawer problem would disappear, and practices that are ethical (but unlawful in the statistical significance regime) such as interim analyses to decide on continuing or stopping data collection, would be acceptable. This change of paradigm suits expertise research. Expertise research suffers from small sample sizes, which are problematic in a world in which individual studies are done to make grandiose claims of statistical significance, but they are perfectly acceptable in a world of slow and humble collective accumulation of evidence. I suggest that pre-registration is a good scientific practice, which should be encouraged when it could be done. However, I criticize the use of pre-registration as a gold standard for publication. In a world without statistical significance, most of the problems that pre-registration aims to solve are no longer problems. I conclude that, for those reasons, retiring statistical significance would boost expertise research.

Keywords

Statistical significance, expertise, evidence, research practices, *p*-value, Bayesian, frequentist

Introduction

In March 2019, the journal *Nature* published the article "Retire statistical significance" authored by Amrhein, Greenland, and McShane (2019), and supported by 854 signatories. I was one of them[1]. In this article I will provide arguments on how retiring statistical significance would benefit expertise research. Let's first examine the problem that retiring statistical significance aims to solve.

The Alleged Problem

Although problems with statistical analysis practices in psychology have a long history (see Cohen, 1994) the article "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant" by Simmons, Nelson, and Simonsohn (2011), initiated a cascade of events that has been changing research practices in psychology.

The article denounced that the researchers' desire to find statistically significant results (i.e., obtaining a *p*-value lower than 0.05) leads them to exercise degrees of freedom in the process of collecting and analysing data. One of the most important problems is conducting interim analyses to decide (a) to continue/stop collecting data; (b) which variables to combine; (c) which variables to exclude/include in the analysis; (d)

which research hypotheses to propose; and (e) to report the analyses that show significant results and not those that are not significant. The consequence of these practices is that an important number of results published in psychological journals are false positives, "the incorrect rejection of a null hypothesis" (Simmons et al., 2011, p. 1359).

The most important research practice that has been proposed to ameliorate this situation is pre-registration (e.g., Nosek, Alter, Banks, Borsboom, Bowman, Breckler, et al., 2015). Pre-registration involves making public research plans, including materials, variables, and hypotheses before starting data collection. After pre-registration the researcher collects data, conducts the pre-registered analyses and submits the corresponding manuscript to an academic journal. The researcher must clearly indicate if there were deviations from the plan, so reviewers and eventual readers can judge whether or not those deviations were justified.

Pre-registration

There is a lot to like about pre-registration. It forces the researcher to plan statistical analyses at the design stage of the study, and it avoids cherry-picking variables and hypotheses after seeing the results. Moreover, pre-registration does not preclude exploratory research; the exploratory researcher only needs to announce that the study is exploratory in the pre-registration material.

However, there are circumstances in which pre-registration is not possible. In the paper "Answering research questions without calculating the mean" (Campitelli, 2015) I described a situation in which one observation can lead to the refutation of a hypothesis. Ericsson, Krampe, and Tesch-Römer (1993) hypothesized that 10 years of intense dedication to a field are *necessary* to achieve high levels of expert performance. Observing one case that violates this rule would refute the hypothesis. An expertise researcher may design a research plan to collect data to test that hypothesis, and in this case could pre-register that plan. However, a researcher who is knowledgeable of archival data on experts' starting ages and achievements before being aware of Ericsson et al.'s hypothesis may already possess the data that refute their hypothesis—the existence of an expert achieving high level of expert performance in less than 10 years of intense practice. Thus, this researcher cannot possibly pre-register a data collection plan because the data were obtained before the researcher was aware of the hypothesis.

If pre-registration is taken as a gold standard for publication, in this case the refutation of an important theory would not be communicated to the scientific community.

Another instance in which pre-registration would not be the gold standard is opportunistic research. Unexpected natural events such as volcano eruptions, floods, and earthquakes are all opportunities for research that do not lend themselves to pre-registration. In expertise research, this may occur when a researcher has an unexpected opportunity to participate in an international/national/local event attended by expert sportspersons, professionals, artists, musicians or scientists. The avid expertise researcher may take advantage of this opportunity to collect data but may not have the time to prepare a detailed research plan. The researcher should not miss the opportunity and should proceed to conduct the study without being penalized for lack of pre-registration, provided the methodology is otherwise sound.

A final objection comes from cognitive modeling. In some studies, using cognitive modeling techniques, adjustments of ancillary, not substantive, aspects of a model are conducted in an iterative fashion. And those adjustments cannot be anticipated before working heavily with the data; hence pre-registration would not be possible, or it would be irrelevant. A discussion of this topic is beyond the scope of this article; for more details see Crüwell, Stefan, and Evans (2019) and the recently published pre-print titled "Preregistration is redundant, at best" by Szollosi, Kellen, Navarro, Shiffrin, van Rooij, Van Zandt, and Donkin (2019).

Redefinition of Statistical Significance

Another solution to the alleged problem, which

can be complemented with pre-registration, was one proposed by 72 authors in the paper "Redefine statistical significance" published in *Nature Human Behaviour* (Benjamin et al., 2017). They asserted: "We propose to change the default *p*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries" (p. 6).

The rationale is the following. A lot of famous effects published in psychology journals come from *p*-values just below the 0.05 threshold for statistical significance. The near threshold effects are the ones that are less likely to replicate. Thus, reducing the threshold to 0.005 will dramatically eliminate false positive results. In my view redefining statistical significance is the wrong solution because it targets the alleged problem, not the real problem.

The Real Problem

The degrees of freedom of the researcher are not the problem; rather, they are the consequence of the real problem. In fact, it is a good thing that researchers exercise degrees of freedom in their research. The problem arises when they exercise degrees of freedom, not to advance scientific knowledge, but to achieve a *p*-value lower than 0.05.

The real problem is, then, that a threshold to determine statistical significance exists. This is compounded by the fact that this threshold is used by editors of academic journals to decide whether a study is worth publishing or not. Editors in academic journals typically reject studies that report an analysis with a *p*-value higher than .05, even when the research design is flawless, or they force the authors to state the result as "negative." This contributes to the "file drawer problem" (i.e., the existence of huge amount of studies with *p*-values higher than .05 sleeping in file drawers and not reported to the scientific community (Rosenthal, 1979)). When researchers attempt to do a meta-analysis to synthesize the knowledge in the field, all the non-significant studies are unlikely to be found; therefore, inflating the size of the effect in the meta-analysis.

The solution to the real problem is surprisingly simple: removing statistical significance, not only as a criterion for publication, but also from statistical analysis. This solution is so simple that I need only to do some clarifications, not an explanation. Removing statistical significance involves eliminating expressions such as "there was a statistically significant difference," "the association of the variables is statistically significant", "the effect of A on B was not statistically significant," etc. This implies the elimination of any threshold to make such decisions, including the more traditional *p*-values of 0.05 or 0.01, the recently "redefined" value of 0.005, or the more recently adopted Bayes factor values and thresholds. That is, the labels suggested by Kass and Raftery (1995) of "not worth more than a mere mention" for Bayes factor values from 1 to 3.2 (or 1 to 1/3.2), substantial evidence for Bayes factor values from 3.2 to 10 (or 1/3.2 to 1/10), strong evidence for values from 10 to 100 (or 1/10 to 1/100), and decisive evidence for values higher than 100 or lower than 1/100 must not be used under this new regime. This does not mean that hypotheses will not be tested or that *p*-values or Bayes factors must not be calculated. It means that, whatever result we find, it will not be classified as statistically significant or not. Let's see in the next section how a scientific world without statistical significance would look.

The World After Retiring Statistical Significance

The proposal of eliminating statistical significance in Amrhein et al. (2019) is not only the opinion of three authors and 854 signatories. The journal *The American Statistician* published an editorial statement of the American Statistical Association (ASA) (Wasserstein & Lazar, 2016) with the title "The ASA's statement on *p*-values: Context, process, and purpose." In that article they warned about misuses of *p*-values. However, ASA fell short on proposing the elimination of statistical significance. That position changed in the editorial article titled "Moving to a world beyond "$p < 0.05$"" published in the same journal by Wasserstein, Schirm, and Lazar

(2019). They asserted that after reviewing 43 articles in that issue and related literature, "it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different", "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way" (p. 2).

What would a world without statistical significance look like? The first important consequence would be that, given that statistical significance is not a publication criterion anymore, the dodgy research practices identified by Simmons et al. (2011) would likely disappear. Consequently, we will see published studies with large effects, medium effects, small effects, negligible effects, and null effects, regardless of whether they would have been significant or not in the previous regime. The standard for publication would be the soundness of the research design, of the tested theories, and of the research questions. Meta-analyses would be more valid because meta-analytic researchers would be able to find the studies that were not published under the previous regime. Thus, the overall effect sizes obtained in meta-analyses will be less biased.

The Results sections of research reports would very much look like the current research reports. Researchers would report descriptive statistics of their samples, they would estimate parameters, and they would compare hypotheses or models.

In parameter estimation, researchers would use traditional methods to obtain a point estimate and interval estimates to show uncertainty in the estimation (or values most compatible with the data, as proposed by Amrheim et al., 2019), as proposed by Cumming's "New Statistics" (Cumming, 2014), if they are frequentists. Alternatively, they can use resampling methods such as bootstrap (Efron & Tibshirani, 1986) to obtain those estimates. If they are Bayesians, they will obtain a posterior distribution of possible parameter values (see Kruschke & Liddell, 2018).

In terms of hypothesis testing, frequentists could still use the NHST approach and come up with a *p*-value. It would be called NHT (null hypothesis testing) rather than NHST (null hypothesis statistical significance testing). They would report "if the null hypothesis of zero effect and all other model assumptions were correct, the probability of obtaining the value of the statistic observed in this study, or a more extreme one, is p". But they will never say "we reject the null hypothesis," "we fail to reject the null hypothesis," "the difference in means is statistically significant," "the correlation failed to reach statistical significance," etc. Of course, there are many other frequentist alternatives, so NHT may disappear altogether.

Bayesian hypothesis testers will compare a model that implements the null hypothesis with a model that implements the alternative hypothesis and come up with a Bayes factor value. If that value is 5 they will say that, given the observed data, the model that implements the alternative hypothesis is 5 times more probable than the model that implements the null hypothesis. A Bayes factor of 0.2 would indicate that the model that implements the null hypothesis is 5 times more probable that the model that implements the alternative hypothesis. A Bayes factor of 1 will indicate that the two models are equally probable. But they will never say that the difference between the models is or is not statistically significant.

Researchers interested in comparing two models, none of which is a null model, will use Akaike Information Criterion (AIC) or Bayes factors to determine which model is relatively more probable, not whether the difference between the models is statistically significant. Notice that the word "relatively" is crucial here: Bayes factors or AICs are not absolute measures. If the compared models are both bad, the Bayes factor in favor of one of the models may be a large number but that model may still be a bad model.

Prohibiting the use of statistical significance may lead to unwanted restrictions. However, it does not seem that the world after it would be too restrictive: It would still accommodate frequentists and Bayesians, parameter estimators, hypothesis testers, and model comparators. Moreover, because statistical significance would not be a criterion for

publication any longer, many incentives for dodgy research practices would disappear.

There is another gain. By retiring statistical significance, some research practices that were unacceptable now become acceptable, or even encouraged. For example, conducting an interim analysis of data, and deciding to stop collecting data or adding more participants based on that analysis, is unacceptable under the statistical significance paradigm. Since participating in research has a degree of discomfort for participants, and it is time and financially consuming, if an interim analysis clearly shows that an expected effect will be very unlikely to be found, it is ethically correct to stop collecting data and report the results. Furthermore, if the interim analysis indicates that we need more data to obtain a result that would provide confidence to researchers, then it would be appropriate to continue collecting data.

When conducting multiple analyses, researchers will be discouraged to hide the "non-significant" ones, for two reasons. First, "non-significant" results will be publishable, so there is no need to hide them. Second, under the new regime, a study with a *p*-value of .049 would look the same whether it is presented on its own or as one of multiple analyses with *p*-values larger than 0.05. Only when the *p*-value is used to make a binary decision of significance would the researcher need to make corrections for multiple comparisons. Without statistical significance the lemma "the data is the data is the data" reigns supreme. The plans of the researcher, other analyses conducted by the researcher (or any other researcher at any point in the history of science, for that matter) are irrelevant for describing the data before us. (They are, of course, relevant for other aspects of research, not for statistical analysis).

Humility

It may seem that sending statistical significance to exile will lead to anarchy: Everything is allowed. Quite the contrary. The law of statistical significance will be replaced with the law of evidence. Conducting statistical analyses will provide researchers with techniques to quantify evidence—evidence for the possible values of parameters, relative evidence for and against hypotheses, relative evidence for and against models. And the language for evidence will be probability:

> If such and such hypothesis and the model assumptions were correct, the probability of observing the summary of the data we observed (or more extreme) is p.
>
> Model A is x times more probable than model B.
>
> In the Bayesian posterior distribution, the range of values from a to b of parameter μ has a probability of p.

A corollary of the new law is that scientists will be forced to exercise humility. They will not be able to claim that an effect was significant; they will only be able to provide a quantity that summarizes the evidence in favor of the effect, or against the lack of effect. Therefore, their claims will be humble. And because humble claims of degree of evidence come with less fanfare than claims of significance, researchers will be more concerned with developing sensible theories and less concerned with finding effects.

What About Decisions?

Statistical significance provides a tool for making decisions. Do chess players have a higher intelligence than non-chess players? P > .05, then probably not. Does this new training method work? P < .05, then yes. The purpose of the first research question is to acquire knowledge about the world, which may or may not have some application in the future. This type of research question perfectly lends itself for the rule of evidence. Researchers do not need to decide whether or not intelligence is associated with chess playing; they are more interested in estimating the size of the difference in intelligence between those groups, with zero difference being a value with no more or less interest than any other value. Also, why waste time making a decision? Science is a cumulative enterprise (see Cumming, 2014), and other researchers will conduct studies that will add information to the current study and collectively determine a degree of evidence. A lot of

research in psychology and in expertise is of the first type; thus, losing a decision-making tool would not cause problems.

The second type of research question may have a direct applicability. If it is found that the new training method has an effect on performance, a coach may adopt that method. Thus, deciding whether the new training method works would be important in this case. However, it would not be sensible to make a decision based solely on the fact that a statistic reached a threshold or not. Rather, decisions should be made on the basis of both the evidence in favor of the new method and the cost of changing to the new method. Another factor to take into account for the decision is the status of the current training method. If the current training method is satisfactory, I would like to see evidence for a large difference between the new method and the current one to adopt the new method. On the other hand, if the current method is giving appalling results, even a small advantage in favor of the new method would be sufficient to decide to adopt it.

The Impact on Expertise Research

In expertise research the observation of unusual events that may refute theories may be more frequent than in other fields (a pirouette that was not considered possible is observed in a gymnast, a world record in athletics or swimming, which was thought impossible to be broken, is broken). The same applies to opportunistic research (an international training camp suddenly occurring in a researcher's hometown, a researcher receiving last minute funding to attend a conference of world class specialists in a field of medicine, etc.).

Moreover, given that high level experts constitute a very small sub-population of the general public, sample sizes are typically small in expertise research. Consequently, expertise researchers would be at a huge disadvantage if pre-registration is taken as a gold-standard criterion for publication. Which non-specialized journal would take seriously a research plan that announces a sample of 20 experts?

But, is a research with that sample size worth it? What can we learn from such a small sample size study? We can learn a little bit and communicate the findings to other researchers. The latter, inspired by this published study, will be able to collect another sample of 20 experts in another part of the world, and so forth, and collectively obtain a decent sample size. This collective effort would not have been possible had the first small sample study not been published. Making a decision regarding statistical significance would be extremely wrong with that small sample. But who needs to make a decision? We rather need to communicate what the current provisional best estimate for a parameter value is, and how precise that estimation is; or which of a few hypotheses is, provisionally, the one that receives more support based on the data.

For those reasons, I believe removing statistical significance would boost expertise research. Expertise researchers who make strenuous efforts to achieve high quality (but low quantity) samples would not feel they are conducting sub-optimal research. Simply because they are not: They are conducting good research, but they are currently forced to use a misleading paradigm that makes their research look bad.

Conclusion

I signed the support for the recommendation of retiring statistical significance from scientific research because I believe it is time statistical practices align with what the goal of science should be: a collective international effort of accumulating evidence and theoretical enrichment, and, consequently, abandoning the egotistic effort of achieving fame by finding statistically significant and gimmicky effects.

Footnote

1. I am also one of the editors of the *Journal of Expertise*. The views in this article, however, are personal.

Author's Declaration

The author declares that there are no personal or financial conflicts of interest regarding the research in this article.

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature, 567*, 305-307.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Cesarini, D. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6-10.

Campitelli, G. (2015). Answering research questions without calculating the mean. *Frontiers in Psychology*, *6*, 1379.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Crüwell, S., Stefan, A. M., & Evans, N. J. (2019). Robust standards in cognitive science. *Computational Brain & Behavior*, 2, 55-65.

Cumming, G. (2014). The New Statistics: Why and how. *Psychological Science, 25*, 7–29.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science, 1*, 54-75.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*, 363-406.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*, 178-206.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Contestabile, M. (2015). Promoting an open research culture. *Science*, *348*, 1422-1425.

Rosenthal, R. (1979). File drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

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.

Szollosi, A., Kellen, D., Navarro, D., Shiffrin, R., van Rooij, I, Van Zandt T, & Donkin, C. (2019). Preregistration is redundant, at best. https://doi.org/10.31234/osf.io/x36pz

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician*, *70*, 129-133.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". *The American Statistician*, *73*, 1-19.

Received: 4 September 2019
Revision received: 18 November 2019
Accepted: 22 November 2019