Dear Prof Lakens,

Thank you and the two reviewers for your very useful comments. I’m happy to hear that you see value in this work. I have addressed each point that was raised both in the manuscript and summarized them below in italics.

Best wishes,

Ian

**Editor**

1.

It’s the scientists who have goals, not psychological science. So, maybe psychological scientists’ ?

*Author response: I have changed this to follow your suggestion.*

2.

Actually, since the algorithm draws from a uniform distribution, it mimics selection bias – doing many tests and picking the significant one. P-hacking is often more strongly associated with small changes in the analysis. The difference is that p-hacking does not lead to a uniform distribution (but the typical ‘bump’ just below 0.05). Something to consider (see other reviewer comment).

*Author response: Thank you for raising this point – I respond to it in a reply to one of R1 point 1 below.*

3

You write: “I refer to this approach as a form of machine learning so as to increase my chance of getting published”. Might be funnier to not make it about you, but write “I recommend researchers to rfer to this this procedure as a form of machine learning to increase their chances of getting published’. Maybe also change “procedure” to “statistical inference procedure”

*Author response: I have changed this to follow your suggestion.*

4

For the R code: probably funnier without the line: “# generate random numbers, stop when < .05” – makes it more subtle.

And the following is not always the output, so I would delete it, and present only the code.

## [1] "p\_ointless = 0.03

*Author response: I have changed this to follow your suggestion.*

5

You write: “Decisions made on the basis of traditional hacked p values and the pointless metric were then compared in a simulation study.” Maybe: “To evaluate the performance of this highly advanced machine learning procedure compare p-hacked p values we performed a simulation study”.

*Author response: I have changed this to follow your suggestion.*

6

You write: “Results demonstrated the results of pointless and traditional p-hacked results are congruent in 100% of cases.” This is funny. An alternative would be: “The p\_ointless procedure performed significantly better (p\_ointless < 0.05).”

*Author response: I sort of like the reflexivity of this joke, but I think it might also be overkill; I’ve kept the original.*

7

The sentence: “As such, this minor discrepancy is easily ignored.” can go.

*Author response: Removed as suggested*

8

You write: “More importantly, execution time for pointless is less than one second, whereas traditional p-hacking techniques can take hours or days.” Maybe add: “in addition to the years of indoctrination needed to get naïve young scholars to believe that “this is how everyone does it”.

*Author response: Changed to “More importantly, execution time for pointless is less than one second, whereas traditional p-hacking techniques can take hours or days to apply – not to mention the years of normalization of p-hacking practices.”*

**Reviewer 1: Neuroskeptic**

1

In this article, the author presents a simple method to "greatly accelerate and streamline the p-hacking process: generating random numbers that are < .05". The method is an R script that repeatedly generates numbers between 0 and 1 and returns the first one to pass the test of being < 0.05.

The article is, of course, tongue in cheek with the point being that p-hacking is equivalent to random number generation until a significant value is obtained. The message is that p-hacked p-values are completely meaningless, so we might as well save time and just generate random numbers instead.

I am sympathetic to this point of view and I found the article amusing. However, I am not sure that "p-hacking is equivalent to infinite random number generation" is true, or at least, I don't think we can assume that it is true.

For a given dataset, it might or might not be possible to obtain a p <0.05 result by p-hacking. I suspect that for the 'average' psychology dataset it would be possible to do this, but this is just an intuition.

To put it another way, I am not sure how many effective degrees of freedom there are in the analysis of the average dataset. Clearly, there are usually thousands if not millions of combinations of analysis choices that one could make, but many of these will give correlated results (similar p-values) because they are minor variations of each other, so the effective degrees of freedom (or number of independent p values we can generate) is lower than one might naively calculate.

It might be that even intense p-hacking of a dataset is only equivalent to the generation of, say, a handful of independent random p-values, in which case there is no guarantee that p<0.05 would be achieved. In that case, p<0.05 results would still be worth something, although less than their "face value", because they would be more likely to occur under H1 than H0.

On the other hand, if the effective degrees of freedom is in the hundreds, at least one p<0.05 is virtually certain, and p<0.05 really would be meaningless. Under this assumption, p\_ointless would indeed be a good model for the average p-hacker's work.

So I think the author needs to, at the least, acknowledge somehow that the whole joke rests on an assumption about the effectiveness of p-hacking. Perhaps they could acknowledge that the "while loop" would not, in fact, run indefinitely in the case of real p-hacking, but that the chance of finding at least one p<0.05 quickly reaches a high value the more cycles are allowed, e.g. the chance is >50% after just 14 cycles (.95^14 = 0.488).

Ideally, the author should then either find a paper in which someone estimated the effective degrees of analytic freedom in the average dataset (it may well have been done), or run such a study themselves. (I suspect that calculating the degrees of freedom would be easy once we know the number of variables and the correlations between them; so it would first be necessary to estimate the average number of variables in a psychology dataset and the collinearity between them.)

*Author response:*

*Thank you for these points – they were useful to chew on.*

*With regard to some of your points as well as Editor comment 2 above, I should note that my choice of implementation of the p\_ointless algorithm here would be better stated as intending to demonstrate its utility via the congruence of its outcomes with traditional p-hacking (i.e., always producing significant results) rather than any claim of correspondence between its implementation and the behaviour of p-hacking a given p value. That is, I’m simply making the point that “a determined p-hacker will always find a significant result”.*

*R2 raises interesting questions here about the limited degrees of freedom in p-hacking. However, I think these comments demonstrate a possible lack of expertise in the dark arts of p-hacking, or possibly has an advanced and debilitating case of research integrity. Poorly justified outlier removal can produce significant results from any dataset, unless the sample size is very low. And, in such cases, the line between removing data points and adding a few is relatively arbitrary. That is, p\_ointless simulates the inevitable dichotomous conclusion that a determined p-hacker will reach - perhaps they would not reach it within every data set, but certainly they can across studies, for example if they are sufficiently convinced of their own hypothesis or have successfully killed the part of their brain that allows them to feel shame.*

*Your useful comments here have allowed me:*

*(1) to refactor the implementation of the p\_ointless algorithm. It is now implemented in 60% fewer lines of code! (2 instead of 5), and*

*(2) to HARK about my original intentions for p\_pointless: I have removed references to the internal workings mapping on to the process of p hacking.*

*“I observed that traditional approaches are relatively time consuming and inefficient (i.e., exploitation of researcher degrees of freedom until p < .05: Simmons et al., 2011). The pointless metric was inspired by the observation that, regardless of the specific p-hacking strategy employed, the product of this process is highlight reliable (i.e., the statistical result “p < .05”).”*

*Whereas the previous implementation sampled from a uniform distribution between 0 and 1 and looped while this value was >.05, this new implementation simply samples from a uniform distribution between 0 and 0.0499. This removes any unintentional implication that there is an intended correspondence between the implementation and any specific form of p-hacking or indeed selection bias (therefore addressing Editor’s point 2).*

*In light of your comments, I gave some thought to whether p\_ointless should sample from a distribution other than uniform in order to, for example, produce a larger proportion of values between 0.02 and 0.05 (as might be observed in traditional hack p values). However, this (a) begins to make it a question of correspondence of implementation with the behaviour rather than congruence of outcomes, and (b) relies on precisely the sort of understanding of p values that this project intends to rid us all of.*

**Reviewer 2: James Heathers**

1

**Satire has a long history and tradition within academic publishing. Often it is met with misunderstanding, which becomes very tiresome. In favour of having to ruin the document by explicitly marking it down as satire, it is much easier to simply make the satire as heavy-handed as possible, which this manages to do very well.**

**My suggestions are as follows:**

"With a few recent and unfortunate exceptions (e.g., Camerer et al., 2018; Klein et al., 2018; Open Science Collaboration, 2015), the discovery that p values can be hacked to support researchers’ hypotheses has proven to be of exceptional utility to the enterprise of psychological science (e.g., acquiring publications, tenure, and flair; see Bakker et al., 2012; Simmons, Nelson, & Simonsohn, 2011 for tutorials)."

**While I don't wish to ruin the brevity of this document, it ruins its utility as a public document to leave 'hacked' and 'flair' undefined. Hacking in particular has some good quotes around it which can be used satirically, and permeates the rest of the article.**

*Author response: I have reworked the opening paragraph to the following:*

*“p-hacking – the updating or adjusting data or analyses in light of prior beliefs about hypotheses – has proven to be of exceptional utility to the goals of psychological scientists (e.g., acquiring high-impact publications, tenure, and paid speaking engagements). While a number of useful tutorials in p-hacking and related strategies exist (e.g., Bakker et al., 2012; Simmons et al., 2011), insightful commentators have pointed out that only those with a ‘flair’ for it are likely to make it in the world of psychological science (Baumeister, 2016). (e.g., acquiring publications, tenure, and flair; see Bakker et al., 2012; Simmons et al., 2011 for tutorials). However, progress has slowed in recent years due to a number of unfortunate setbacks, including wider use of replication and pre-registration (e.g., Munafò et al., 2017; Open Science Collaboration, 2015) by methodological terrorists (Fiske, 2016) and data parasites (Longo & Drazen, 2016).”*

2  
"However, efforts to further optimize the process of p-hacking have slowed in recent years due to a number of unfortunate setbacks such as wider use of replication and pre-registration (Munafò et al., 2017; Nosek et al., 2015; Nosek, Ebersole, DeHaven, & Mellor, 2018).

In this article, I introduce the pointlessmetric and demonstrate how it can streamline the process of p-hacking your results. While this metric does suffer from the mild flaw of providing zero diagnosticity of the presence or absence of a true effect, this property is largely irrelevant to psychological science’s primary goals (e.g., high impact publications and tenure)."  
**Include a sop to the ostensive goals of occasionally producing repeatable definitions of human behaviour as an afterthought.**

*Author response: Added “Secondary goals such as valid and useful insights into human behaviour are also occasionally met, albeit incidentally.”*

3

The pointlessmetric was inspired by the observation that, regardless of the specifics of any given p-hacking strategy, the product of this process is highlight reliable (p < .05)."  
**>Explain briefly.**

*Author response: I have elaborated that the p value I report is not the outcome of test but a reference to this universal outcome of p-hacking:*

*“The pointless metric was inspired by the observation that, regardless of the specific p-hacking strategy employed, the product of this process is highlight reliable (i.e., the statistical result “p < .05”).”*

4  
I refer to this approach as a form of machine learning so as to increase my chance of getting published."

**>Include reason of 'ostensible sophistication'**

*Author response: I’ve left this point unchanged so as to not labour the point (although perhaps I misunderstood the suggestion here).*

5

Decisions made on the basis of traditional hacked p values and the pointlessmetric were then compared in a simulation study. In line with modal p-hacking practices, only the key property of diagnosticity for publishability (i.e., p < .05) was considered. 10,000 cases were simulated (see Appendix for R code). Results demonstrated the results of pointlessand traditional p-hacked results are congruent in 100% of cases. Although variation in individual coefficients frequently differ by large margins, both strategies satisfy the core criterion of producing significant results. As such, this minor discrepancy is easily ignored. More importantly, execution time for pointlessis less than one second, whereas traditional p-hacking techniques can take hours or days."

**>Or alternatively an entire experimental series which use a great deal of public money.**

*Author response: I agree and indeed this point was originally included in the discussion – I removed as I thought it was too hot a take. I’ve added a future directions paragraph to the end of the discussion:*

*“Now that the data processing and analytic process has been streamlined, future work should consider whether data collection itself may be redundant or an inefficient use of researchers’ time. A pilot study by Prof Diederik Stapel suggests that primary goals (e.g., tenure) can indeed be achieved without it (Verfaellie & McGwin, 2011).”*