Dear Dr Hussey,

Thank you for the submission of your manuscript detailing a highly innovative analysis strategy to Meta-Psychology. We believe the analysis strategy you propose has a solid justification, and you realistically argue for an increase in productivity of all psychologists. The reduction in time psychology students will be required to follow statistics training will be greatly appreciated around the world. I was very enthusiastic about this manuscript, as were the reviewers, both experts in p-hacking and satire. The reviewers (Neuroskeptic and James Heathers) provide some useful suggestions for improvement, and I have added some suggestions myself. All are for your consideration. I do not expect to have to send your revised manuscript out for review after you have made any minor revisions based on this feedback.

The summer holiday, and eventful lives of reviewers, lead to some delay – I hope for your understanding.

Best,

Daniel Lakens

My own minor suggestions:

this property is largely irrelevant to psychological science’s primary goals (e.g., high impact publications and tenure). > It’s the scientists who have goals, not psychological science. So, maybe psychological scientists’ ?

You write: “Loosely speaking, this algorithm mimics the outcomes of overt p-hacking behaviour” 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).

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 refer 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”

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

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”.

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).”

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

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”.

Reviewer 1: Neuroskeptic

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.)

Reviewer 2: James Heathers

**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.**  
  
"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.**  
  
"More importantly, the metric possesses three superior characteristics. First, it is non-inferior to current p-hacking practices, which also tell us little about the presence or absence of a true effect (large scale replications put this diagnosticity at no better than a coin toss: Klein et al., 2018). Second, it retains a far more important property of hacked p values: by guaranteeing significant results, it has higher predictive validity for publishability. Finally, it also provides economic benefits relative to the high total life-cycle costs associated with traditional p-hacking (e.g., by eliminating the need for comprehensive graduate training in statistics).

Methods and results

The pointlessmetric follows the same internal logic as traditional p-hacked analytic strategies (e.g., Bem, 2011). Loosely speaking, this algorithm mimics the outcomes of overt p-hacking behaviour (i.e., exploitation of researcher degrees of freedom until p < .05: Simmons et al., 2011). 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.**  
  
"As such, many intermediary steps are therefore arguably unnecessary, and the same end result can be obtained more efficiently by automation. This is accomplished by generating random numbers until one is found that is < .05. 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'**  
  
"R code to calculate pointlessis provided below:

# generate random numbers, stop when < .05

p\_ointless <- 1  
**while** (p\_ointless **>=** .05) {  
  p\_ointless <- **runif**(n = 1)  
}  
  
# print this value  
**print**(**paste**("p\_ointless =", **round**(p\_ointless, 3)))

## [1] "p\_ointless = 0.032"

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.**

Discussion

Traditional p-hacking involves starting with a sound analytic strategy and then iteratively degrading this until the results support one’s hypothesis. On the basis that this strategy almost invariably returns significant results, many burdensome aspects of this analytic process can arguably be bypassed via automation. The most parsimonious method was selected: random number generation. Results from a simulation study demonstrate that decision making on the basis of traditional hacked p values and pointlessare equivalent, and that the latter requires several orders of magnitude less time and resources to calculate. Academic productivity and more importantly flair can therefore be greatly increased through the widespread adoption of this approach.