RUNNING HEAD: WHAT *P*-HACKING REALLY LOOKS LIKE

What *p*-hacking really looks like: A comment on Masicampo & Lalande (2012)

Daniël Lakens

Eindhoven University of Technology

Word Count: 1015 (without references and figure captions)

In Press, Quarterly Journal of Experimental Psychology

Keywords: Statistics; Statistical inference; Hypothesis testing.

Author Note: I'd like to thank E.J. Masicampo and Daniel Lalande for sharing their data, as well as providing answers to my questions. Thanks to Ryne Sherman for his *p*-hack code in R, and Nick Brown and Tal Yarkoni for comments and suggestions. All files required to reproduce the data in this article are available from https://osf.io/ycag9/

Correspondence can be addressed to Daniël Lakens, Human Technology Interaction Group, IPO 1.33, PO Box 513, 5600MB Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl.

Masicampo and Lalande (2012; M&L) assessed the distribution of 3627 exactly calculated *p*-values between 0.01 and 0.10 from 12 issues of three journals. The authors concluded that "The number of *p*-values in the psychology literature that barely meet the criterion for statistical significance (i.e., that fall just below .05) is unusually large". "Specifically, the number of *p*-values between .045 and .050 was higher than that predicted based on the overall distribution of *p*."

There are four factors that determine the distribution of *p*-values, namely the number of studies examining true effect and false effects, the power of the studies that examine true effects, the frequency of Type 1 error rates (and how they were inflated), and publication bias. Due to publication bias, we should expect a substantial drop in the frequency with which *p*-values above .05 appear in the literature. True effects yield a right-skewed *p*-curve (the higher the power, the steeper the curve, e.g., Sellke, Bayarri, & Berger, 2001). When the null-hypothesis is true the *p*-curve is uniformly distributed, but when the Type 1 error rate is inflated due to flexibility in the data-analysis, the *p*-curve could become left-skewed below *p*-values of .05.

M&L (and others, e.g., Leggett, Thomas, Loetscher, & Nicholls, 2013) model *p*-values based on a single exponential curve estimation procedure that provides the best fit of *p*-values between .01 and .10 (see Figure 3, right pane). This is not a valid approach because *p*-values above and below *p*=.05 do not lie on a continuous curve due to publication bias. It is therefore not surprising, nor indicative of a prevalence of *p*-values just below .05, that their single curve doesn't fit the data very well, nor that Chi-squared tests show the residuals (especially those just below .05) are not randomly distributed.

P-hacking does not create a peak in *p*-values just below .05. Actually, *p*-hacking does not even have to lead to a left-skewed *p*-curve. If you perform multiple independent tests in a study where the null-hypothesis is true the Type one error rate is substantially increased, but

the *p*-curve is uniform, as if you had performed 5 independent studies. The right skew (in addition to the overall increase in false positives) emerges through dependencies in the data in a repeated testing procedure, such as collecting data, performing a test, collecting additional data, and analyzing the old and new data together. In Figure 1 two multiple testing scenarios (comparing a single mean to up to 5 other means, or collecting additional participants up to a maximum of five times) are simulated 100000 times when there is no true effect (for details, see the supplementary material). Only 500 significant Type 1 errors should be observed in each bin without *p*-hacking, but we see an increase in false positives above 500 for most of the 10 bins.

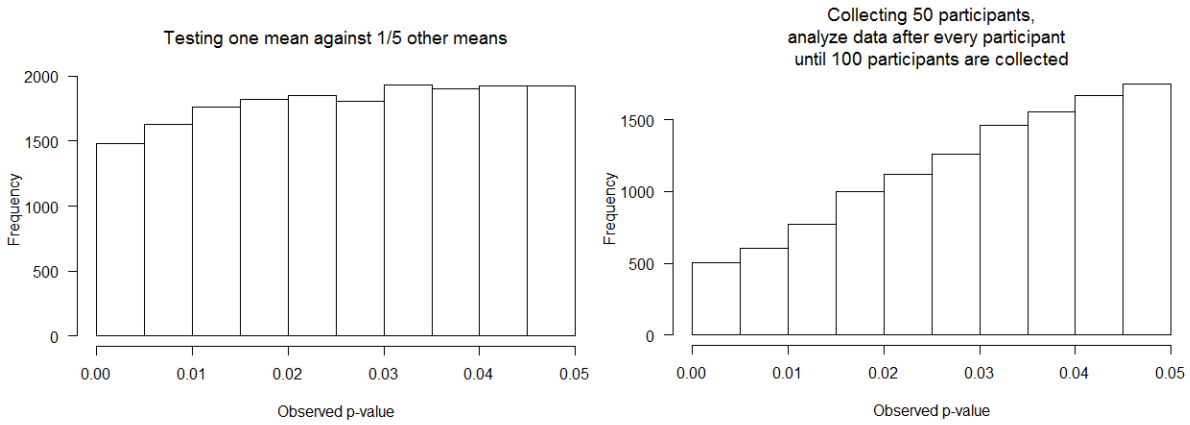

Figure 1. *P*-curves under two scenarios of flexibility in the data analysis (for details, see supplementary material).

Identifying a prevalence of Type 1 errors in a large heterogeneous set of studies is, regrettably, even more problematic due to the *p*-curve of true effects. In Figure 2 (left) we see a *p*-curve of 100000 experiments with 50% power (for details, see supplementary material). Adding the 200000 experiments simulated above gives the *p*-curve on the right. Even when only 1/3rd of the studies examines a true effect with a meager 50% power, it is already impossible to observe a strong left-skewed distribution.

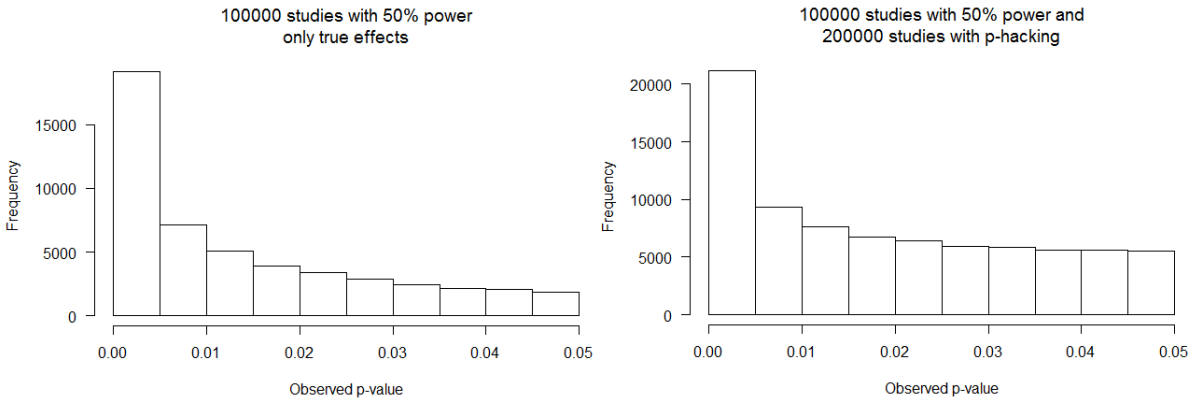

Figure 2. 100000 studies examining true effect with 50% power (left), supplemented with 200000 *p*-hacked studies (right).

Do frequencies of *p*-values just below *p*=.05 observed by M&L indicate extreme *p*-hacking in a field almost devoid of true effects? No. The striking illustration of the prevalence of *p*-values in Psychological Science just below .05 (Figure 3, right, from M&L) from the blind rater is not apparent in the data coded (but not presented individually) by the authors themselves (Figure 3, left, re-analysis based on the data kindly provided by M&L). Similarly, the peak just below *p*=.05 observed in 2005 issues of JPSP coded by Leggett et al., (2013) and attributed to an increase in 'just significant' *p*-values over time does not replicate in the *p*-values M&L collected from 2007-2008 JPSP issues (see supplementary material). Clearly, more data is needed, and the reliability and reproducibility of the analysis of *p*-curves can be improved by always publishing a *p*-curve disclosure table (see Simonsohn, Nelson, & Simmons, 2014).

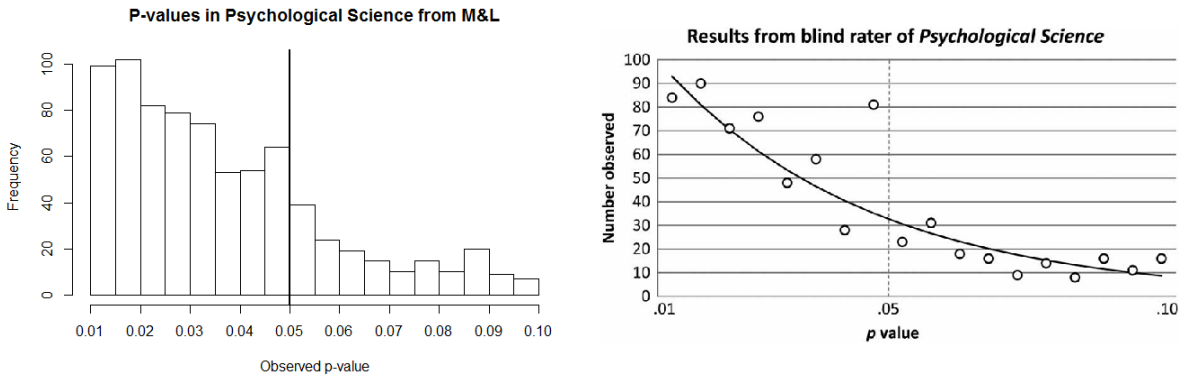

Figure 3. Re-analyzed *p*-curve for *p*-values from Psychological Science coded by M&L (left, reported together with *p*-values from JPSP and JEPG in Figure 1), and the *p*-curve for *p*-values from Psychological Science coded by a blind rater, as reported in M&L Figure 2 (right).

There is also no evidence of a pre-valence of *p*-values just below .05 when analyzing all *p*-values collected by M&L. The authors find no peak when dividing the *p*-values in bins of .01, .005, or .0025, and there is only a slight increase in the .04875-.05 range, which might simply be random variation. Figure 4 (left) presents the outcome of a model of the *p*-curve based on power, publication bias, the Type 1 errors, and the relative frequency of studies examining true and false hypotheses (for details, see the supplementary materials). More research is needed to examine the most probable values for these parameters for specific research areas, but Figure 4 illustrates that these four parameters can in principle quite accurately simulate the *p*-curve observed by M&L based on all coded *p*-values (Figure 4, right). Based on this model, there seems to be a slight increase of *p*-values between .050–0.055 - perhaps reflecting leniency towards studies that are almost statistically significant.

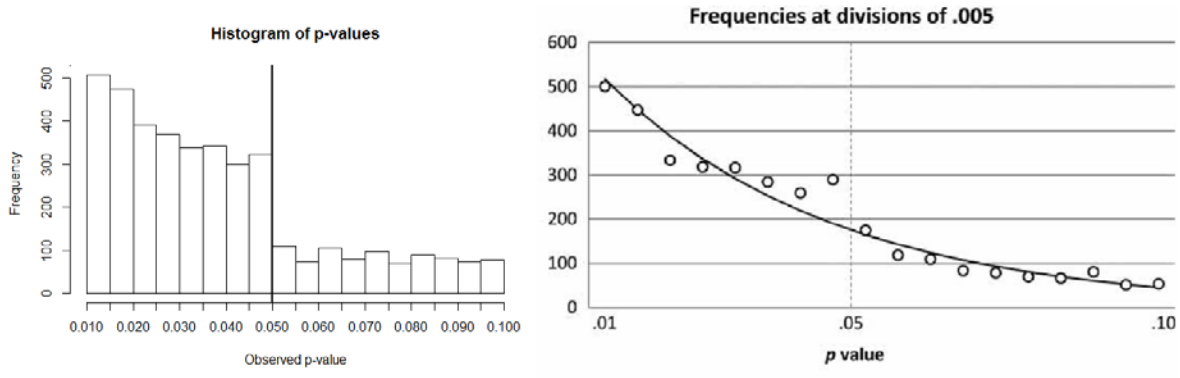

Figure 4. Simulated *p*-values (11000 studies with 41% power, removing half of the studies with a *p*>.05, and adding 844 false positives (left, see supplementary material) and all observed *p*-values by M&L (right)

Even though 844 of the 3907 *p*-values in Figure 4 are Type 1 errors (616 of which are only significant through *p*-hacking) there is no noticeable prevalence of *p*-values just below .05. Altogether, the evidence for a reliable peak of *p*-values just below *p*=.05 in the data collected by M&L is weak. Furthermore, looking for such a peak distracts from the fact that that *p*-hacking will lead to a much greater absolute increase in false positives between 0.01-0.045 than between 0.045-0.05. It should be clear that *p*-hacking can be a big problem even when it is difficult to observe. Although the data by M&L do not indicate a surprising prevalence of *p*-values just below .05 when interpreted against a more realistic model of expected *p*-curves, there is a clear drop in expected *p*-values above .05, which is in line with the strong effect of publication bias on which *p*-values end up in the literature (see also Kühberger, Fritz, & Scherndl, 2014).

An alternative to attempting to point out *p*-hacking in the entire psychological literature is to identify left-skewed *p*-curves in small sets of more heterogeneous studies (i.e., where all studies examine a null-hypothesis that is true). Better yet, we should aim to control the Type 1 error rate for the findings reported in an article. Pre-registration and/or replication (e.g., Nosek & Lakens, 2014) are two approaches that can improve the reliability of findings.

References

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE 9(9)*: *e105825*. doi:10.1371/journal.pone.0105825

Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. E. (2013). The life of *p*: "Just significant" results are on the rise. *The Quarterly Journal of Experimental Psychology*, *66*, 2303-2309. doi: 10.1080/17470218.2013.863371

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

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *43*, 137-141. DOI: 10.1027/1864-9335/a000192.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of *p*-values for testing precise null hypotheses. *The American Statistician*, *55*, 62–71.

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. doi: 10.1037/a0033242.