A peculiar surge of incorrect conclusions about the prevalence of *p*-values just below .05

Daniel Lakens

De Winter and Dodou (2015) analyzed the distribution (and its change over time) of a large number of *p*-values automatically extracted from abstracts in the scientific literature. They concluded there is a ‘surge of *p*-values between 0.041-0.049 in recent decades’ which 'suggests (but does not prove) questionable research practices have increased over the past 25 years'. I show the changes in the ratios of *p*-values over the years between 0.041-0.049 are better explained by a model of *p*-value distributions that assumes the average power has decreased over time. Furthermore, I propose that their observation that *p*-values just below 0.05 increase more strongly than *p*-values above 0.05 can be explained by an increase in publication bias over the years (cf. Fanelli, 2012), which has led to a relative decrease of 'marginally significant' *p*-values in the literature (instead of an increase in *p*-values just below 0.05). I explain why researchers analyzing large numbers of *p*-values in the scientific literature need to develop better models of *p*-value distributions before drawing conclusion about questionable research practices. These analyses highlight that publication bias and underpowered studies are a much bigger problem for science than inflated Type 1 error rates.

RUNNING HEAD: PREVALENCE OF P-VALUES

A peculiar surge of incorrect conclusions about the prevalence of *p*-values just below .05

Daniël Lakens

Eindhoven University of Technology

Word Count: 4408

Author Note: I want to thank De Winter and Dodou for sharing their data, assisting in the re-analysis, and reading an earlier version of this draft. All files required to reproduce the analyses in this article are available from https://osf.io/ms4x6/

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.

Abstract

De Winter and Dodou (2015) analyzed the distribution (and its change over time) of a large number of *p*-values automatically extracted from abstracts in the scientific literature. They concluded there is a 'surge of *p*-values between 0.041-0.049 in recent decades' which 'suggests (but does not prove) questionable research practices have increased over the past 25 years'. I show the changes in the ratios of *p*-values over the years between 0.041-0.049 are better explained by a model of *p*-value distributions that assumes the average power has decreased over time. Furthermore, I propose that their observation that *p*-values just below 0.05 increase more strongly than *p*-values above 0.05 can be explained by an increase in publication bias over the years (cf. Fanelli, 2012), which has led to a relative decrease of 'marginally significant' *p*-values in the literature (instead of an increase in *p*-values just below 0.05). I explain why researchers analyzing large numbers of *p*-values in the scientific literature need to develop better models of *p*-value distributions before drawing conclusion about questionable research practices. These analyses highlight that publication bias and underpowered studies are a much bigger problem for science than inflated Type 1 error rates.

A peculiar surge of incorrect conclusions about the prevalence of *p*-values just below .05

In recent years researchers have become more aware of how flexibility during the data-analysis can increase false positive results (e.g., Simmons, Nelson, & Simonsohn, 2011). If the true Type 1 error rate is substantially inflated because researchers analyze their data until a *p*-value smaller than 0.05 is observed this can substantially decrease the robustness of scientific knowledge. However, as Stroebe and Strack (2014, p. 60) have pointed out: "*Thus far, however, no solid data exist on the prevalence of such research practices*". Some researchers have attempted to provide an indication of the prevalence of questionable research practices by analyzing the distribution of *p*-values in the published literature. The idea is that questionable research practices lead to 'a peculiar prevalence of *p*-values just below 0.05' (Masicampo & Lalande, 2012) or the observation that '"just significant" results are on the rise' (Leggett, Loetscher, & Nichols, 2013).

Despite the attention grabbing titles of these publications, the reported data does not afford the strong conclusions these researchers have drawn. The observed pattern of a peak of *p*-values just below 0.05 in Leggett et al (2013) does not replicate in other collected *p*-value distributions for the same journal in later years (Masicampo & Lalande, 2012), in psychology in general (Kühberger, Fritz, & Scherndl, 2014), or in scientific journals in general (De Winter & Dodou, 2015). The peak in *p*-values observed in Masicampo & Lalande (2012) is only surprising compared to an incorrectly modelled *p*-value distribution that ignores publication bias and its effect on the frequency of *p*-values above 0.05 (Lakens, 2014a).

Recently, De Winter and Dodou (2015) have contributed to this emerging literature on *p*-value distributions and concluded that there is a 'surge of *p*-values between 0.041-0.049 in recent decades'. They improved upon earlier approaches to analyze *p*-value distributions by comparing the percentage of *p*-values over time (from 1990-2013). Two observations in the data they collected could seduce researchers to draw conclusions about a rise of *p*-values just below a significance level of 0.05. The first observation is a much stronger rise in *p*-values between 0.041 and 0.049 than in *p*-values between 0.051-0.059. The second observation is that the percentage of *p*-values that falls between 0.041-0.049 has increased more from 1990 to 2013 than the increase in the percentage of *p*-values between 0.01-0.09, 0.011-0.019, 0.021-0.029, and 0.031-0.039 over the same years[1]. The authors (2015, p. 37)

71 conclude that: "The fact that *p*-values just below 0.05 exhibited the fastest increase among all *p*-value
72 ranges we searched for suggests (but does not prove) that questionable research practices have increased
73 over the past 25 years."

74 I will explain why the data does not provide any indication of an increase in questionable
75 research practices. First, I will discuss how the difference in the increase in *p*-values just below 0.05
76 and just above 0.05 is due to publication bias, where (perhaps surprisingly) *p*-values just above 0.05 are
77 becoming relatively less likely to appear in the abstracts of published articles over the years. Second, I
78 will explain why the relatively high increase in *p*-values between 0.041-0.049 over the years can easily
79 be accounted for by a decrease in the average power of studies, but is unlikely to emerge due to an
80 inflated Type 1 error rate due to questionable research practices. I want to explicitly note that it was
81 possible to provide these alternative interpretations of the data mainly because the authors shared all
82 data and analysis scripts online. While I criticize their interpretation of data, I applaud their adherence
83 to open science principles which greatly facilitated cumulative science.

84 As I have discussed before (Lakens, 2014a), it is essential to accurately model *p*-value
85 distributions before drawing conclusions about *p*-values extracted from the scientific literature.
86 Statements about *p*-value distributions require a definition of four parameters. First, researchers should
87 specify the number of studies where H0 is true, and the number of studies where H1 is true. Second,
88 researchers need to estimate the average power of the studies (or the average power of multiple subsets
89 of studies, if heterogeneity in power is substantial). Third, the true Type 1 error rate and any possible
90 mechanisms through which the error rate is inflated should be specified. And finally, publication bias,
91 and a model of how the *p*-value distribution is affected by publication bias, should be proposed. It is
92 important to look beyond simplistic comparisons between *p*-values just below 0.05 and *p*-values in
93 other locations in the *p*-value distribution outside the scope of an explicit model of the four parameters
94 that determine *p*-value distributions.

95 **Are *p*-values below 0.05 increasing, or *p*-values above 0.05 decreasing?**

96 De Winter and Dodou (2015) show there is a relatively stronger increase in *p*-values between
97 0.041-0.049 than between 0.051-0.059 (see for example their Figure 9). The data is clear, but the reason
98 for this difference is not, and it is not explored by the authors. Are *p*-values below 0.05 increasing more,

or are *p*-values above 0.05 increasing less? A direct comparison is difficult, because the percentage of papers reporting *p*-values below 0.05 can increase due to an increase in *p*-hacking, but also due to an increase in publication bias. If publication bias increases, and people report less non-significant results, the percentage of papers reporting *p*-values smaller than 0.05 will also increase, even if there is no increase in *p*-hacking. Indeed, Fanelli (2012) has shown negative results have been disappearing from the literature between 1990-2007, which would explain the relative differences in *p*-values between 0.041-0.049 and 0.051-0.059 observed by De Winter and Dodou (2015).

We can examine the alternative explanation that the relative differences observed are due to publication bias increasing, instead of due to an increase in *p*-hacking, by comparing the relative differences between *p*-values between 0.031-0.039 and 0.041-0.049 over the years on the one hand, and 0.051-0.059 and 0.061-0.069 on the other hand. If there is an increase in *p*-hacking, the biggest differences should be observed below 0.05 (in line with the idea of a surge of *p*-values between 0.041-0.049)[2]. However, there are reasons to assume the biggest difference might occur in *p*-values just above 0.05. As Lakens (2014a) noted, there seems to be some tolerance for *p*-values just above 0.05 to be published, as indicated by a higher prevalence of *p*-values between 0.051-0.059 than would be expected based on the power of statistical tests and an equal reduction of all *p*-values above 0.05. If publication bias becomes more severe, we might expect a reduction in the tolerance for *p*-values just above 0.05, which would lead to the largest changes in ratios above 0.05.

Across the three time periods (1990-1997, 1998-2005, and 2006-2013) the ratio of *p*-values in the 0.03 range to *p*-values in the 0.04 range is pretty stable: 1.13, 1.09, and 1.11, respectively. The ratio of *p*-values in the 0.05 range shows a surprisingly large reduction over the years: 2.27, 1.94, and 1.79, respectively. This surprisingly large change in ratios over time for *p*-values between 0.051-0.059 indicates that instead of a surge of *p*-hacking, the real change over time happens in the *p*-values between 0.051-0.059, which is not in line with an explanation based on an increase in questionable research practices over time.

This might be explained by the idea that where *p*-values between .051-0.59 (or 'marginally significant' results) were more readily interpreted as support for the hypothesis in 1990-1997 than in 2005-2013. This idea is speculative, but seems likely given the increase in publication bias over the

years (Fanelli, 2012). It should be noted that *p*-values just above the 0.05 level are *still* more frequent than can be explained just by the average power of the tests and publication bias that is equal for all *p*-values above 0.05 (cf. Lakens, 2014a). In other words, this data is in line with the idea that publication bias is still slightly less severe for *p*-values just above 0.05, even though this benefit of *p*-values just above 0.05 has become smaller over the years.

To summarize, De Winter & Dodou (2015) show a relative difference in the increase in *p*-values just above 0.05 and just below 0.05, but do not examine the possible reasons for this difference. I show that the strongest difference in ratios for *p*-values above 0.01 occurs for *p*-values between 0.051-0059, which seems to be the driving force for the differences in the increase of *p*-values in the 0.041-0.049 range and *p*-values in the 0.051-0.059 range reported by De Winter and Dodou (2015, e.g., Figures 9 and 10). These observed differences provide no indication for a surge of *p*-values between 0.041-0.049 over the years due to an increase in questionable research practices, but instead require an explanation for the surprising relatively smaller increase in *p*-values between 0.051-0.059. Since previous research has revealed there is an increase in publication bias over the years (Fanelli, 2012), one possible mechanism for the relatively smaller increase of *p*-values between 0.051-0.059 compared to *p*-values between 0.041-0.049 is an increase in publication bias.

How changes in average power over the years affect ratios of *p*-values below 0.05

The first part of the title of the article by De Winter and Dodou (2015), "A surge of *p*-values between 0.041-0.049" is based on the observation that the ratio of *p*-values between 0.041-0.049 increases more than the ratio of *p*-values between 0.031-0.039, 0.021-0.029, and 0.011-0.019. There are no statistics reported to indicate whether these differences in ratios are actually statistically significant, nor are effect sizes reported to indicate whether the differences are practically significant (or justify the term 'surge'), but the ratios do increase as you move from bins of low *p*-values between 0.001-0.009 to bins of high *p*-values between 0.041-0.049.

The first thing to understand is why none of the observed ratios are even close to 1. The reason is that there is a massive increase in the percentage of papers in which *p*-values are reported over the years. As De Winter & Dodou (2015, p. 15) note: "*In 1990, 0.019% of papers (106 out of 563,023 papers) reported a p-value between 0.051 and 0.059. This increased 3.6-fold to 0.067% (1,549 out of*

2,317,062 papers) in 2013. Positive results increased 10.3-fold in the same period: from 0.030% (171 out of 563,023 papers) in 1990 to 0.314% (7,266 out of 2,317,062 papers) in 2013." This is not just an increase in the absolute number of reported *p*-values in abstracts (in which case the ratios could still be 1) but a relative 10.3-fold increase in how often *p*-values end up in abstracts. De Winter and Dodou (2015) show *p*-values are finding their way into more and more abstracts, which points to a possible increase in the overreliance on null-hypothesis testing in empirical articles. This is an important contribution to the literature, even when other claims about an increase in questionable research practices would not hold.

The main question is how these differences in the ratios across the 5 bins below 0.05 be explained. De Winter and Dodou (2015) do not attempt to provide a model that explains the observed *p*-value distribution, but mathematically, any model of *p*-value distributions needs to specify the ratio of true to false effects examined, the average power of the studies, the Type 1 error rate, and publication bias. It is only possible to explain the relative differences between the ratios of the different bins of *p*-values if we allow at least one of the parameters of the model to the change over time. Because we are focusing on the *p*-values below 0.05 we can ignore publication bias, assuming all disciplines that report *p*-values in abstracts use $\alpha = 0.05$ (this is not true, but we can assume it applies to the majority of articles that are analyzed). The two remaining possibilities are a change in the average power of studies over time, and an inflated Type 1 error rate over time (such as an increase in questionable research practices in the literature).

If we ignore Type 1 errors, we can relatively easily reconstruct the observed data purely based on differences in the average power across the years. I'm not arguing the numbers in this re-construction reflect the truth. However, any model of the *p*-value distribution *must* estimate the average power of the studies. It is not difficult to model the ratios observed by De Winter & Dodou (2015) under the assumption that power decreases from 1990 to 2013. For example, if we assume the average power of studies was 55% in 1990, and 42% in 2013, we can expect to observe the *p*-value distribution across the 5 bins as detailed in the table below, with 29.86% of the *p*-values falling between 0.001 and 0.009 in 1990, but only 19.93% of *p*-values falling between 0.001-0.009 in 2013 (which most likely explains

the large differences in ratios between 0.001-0.009 discussed earlier). This is just the *p*-value distribution as a function of the power of the tests.

Table 1: Expected percentage of *p*-values between 0.001-0.049 based on 42% and 55% power.

	55% power	42% power
p0.001-p0.009	29.86	19.93
p0.011-p0.019	8.54	7.22
p0.021-p0.029	5.61	5.06
p0.031-p0.039	4.22	3.98
p0.041-p0.049	3.39	3.37

If we incorporate the fact that the percentage of *p*-values reported in the abstract has increased by 10% over the years (column 2 and 3 in Table 2 below), and use as total studies in 1990 563023, and as total studies in 2013 2317062 (taken from De Winter & Dodou, 2015) then we should expect the total number of observed *p*-values in 1990 and 2013 to approximate those displayed in the reconstructed # of p-values columns below. By choosing an average power of 55% for studies in 1990, and an average power of 42 in 2013, these numbers mirror the observed # of p-values by De Winter and Dodou (2015).

Table 2. Percentage of papers that report *p*-values in abstracts, and the number of reconstructed and observed (De Winter & Dodou, 2015) *p*-values between 0.001-0.049 in 1990 and 2013.

	% *p*-values in abstracts	% *p*-values in abstracts	reconstructed # *p*-values 1990	reconstructed # *p*-values 2013	observed # *p*-values 1990	observed # *p*-values 2013
p0.001-p0.009	0.01	0.1	1681	46170	1770	44970
p0.011-p0.019	0.01	0.1	481	16728	462	14885
p0.021-p0.029	0.01	0.1	316	11725	268	10630
p0.031-p0.039	0.01	0.1	238	9210	240	9108
p0.041-p0.049	0.01	0.1	191	7646	178	8250

When we calculate the ratios of the observed *p*-values, we see in Table 3 they approach the general pattern of the ratios observed by De Winter and Dodou (2015). The reconstruction is not perfect, for a number of reasons. First of all, there is very little data from 1990, which will lead to substantial variation between expected and observed frequencies for any model (the fit of the model increases for comparisons between years where there is more data available). For example, the fact that the difference

in the percentage of *p*-values in the 0.021-0.029 bin from 1990 to 2013 is larger than for *p*-values in the 0.031-0.039 bin is only true in 1990 and 2008, but is reversed (as predicted by a model of *p*-value distributions where power changes over time) in the remaining 21 comparisons of 2013 with each preceding year.

Table 3. Ratios of reconstructed *p*-values and *p*-value ratios observed by De Winter & Dodou (2015) between 0.001-0.049 for 1990 and 2013.

	reconstructed ratio N/T 1990	reconstructed ratio N/T 2013	reconstructed 1990/2013 Ratio	observed ratio N/T 1990	observed ratio N/T 2013	observed 1990/2013 Ratio
p0.001-p0.009	0.306	1.993	6.674	0.315	1.945	6.17
p0.011-p0.019	0.085	0.722	8.454	0.082	0.644	7.83
p0.021-p0.029	0.056	0.506	9.017	0.048	0.460	9.63
p0.031-p0.039	0.042	0.398	9.417	0.043	0.394	9.21
p0.041-p0.049	0.034	0.330	9.740	0.032	0.367	11.28

Similarly, when comparing 2013 to each of the 23 preceding years, the ratio is higher for *p*-values between 0.041-0.049 than for 0.031-0.039 in 12 out of 23 comparisons – only just more than 50% of the time, which can hardly be called a 'surge' of *p*-values between 0.041-0.049. The model based on power differences predicts that ratios for *p*-values between 0.031-0.039 should be very similar to those between 0.041-0.049. Given the small percentages of articles that report *p*-values and the variation inherent in observed *p*-value distributions, it is not surprising the ratios for 0.041-0.049 are only just more than 50% likely to be higher than those for *p*-values between 0.031-0.039. This observation is more difficult to explain based on the idea that questionable research practices have increased, which typically assumes *p*-values between 0.041-0.049 increase more strongly than *p*-values between 0.031-0.039 (e.g., Leggett et al., 2013; Masicampo & Lalande, 2012).

Obviously this model is too simplistic. It does not include any Type 1 errors, and it assumes homogeneity in the power of the performed tests. We can be certain power varies substantially across studies and research disciplines, and we can be certain the assumption that the *p*-value distribution perfectly follows the distribution based on a single average power value. The *p*-value distribution can only be reconstructed exactly if we know how many studies had which specific power, but this is

impossible. For the current purpose, which is to demonstrate the observed pattern can be reconstructed by assuming the average power has changed over time, a more advanced model is not required. However, future attempts to provide support for an increase in Type 1 errors, or attempts to calculate average effect sizes based on *p*-value distributions (e.g., van Assen, van Aert, & Wicherts, in press; Simonsohn, Nelson, & Simmons, 2014) need to develop more detailed models of *p*-value distributions. For now, the most important conclusion is that a change in power over time can mathematically account for the observed changes in ratios in the different *p*-value bins. Moreover, the idea that power decreases over time is theoretically plausible, since such a decrease in power over time has been observed in some disciplines, such as psychology (Sedlmeier & Gigerenzer, 1989).

Let's assume the average power has not changed over time, and instead try to reconstruct the observed ratios by a change in the Type 1 error rates over time. As long as the Type 1 error rates are the same for each bin of *p*-values, the ratios equal the overall increase in *p*-values reported in abstracts over time. To reconstruct the ratios as observed by De Winter and Dodou (2015), we need to assume *p*-hacking leads to a stronger increase in higher *p*-values than in lower *p*-values. Although this is a reasonable assumption under many types of *p*-hacking, it turns out to that the specific pattern of inflated Type 1 error rates required to reconstruct the observed ratios in not very likely to emerge in real life.

To simulate the impact of questionable research practices, we need to decide upon the ratio of studies where H0 is true and studies where H1 is true, and the exact increase in Type 1 error rates for each bin of *p*-values below 0.05. Type 1 errors come exclusively from analyzing results of studies where H0 is true (*p*-hacking when H1 is true inflates the effect size estimate, and thus can be seen as an incorrect way to increase the power of a test which leads to an overestimation of effect sizes). In the calculations below, power is kept constant, but inflated Type 1 error rates are introduced. This is the equivalent of the true power of studies reducing over the years, which is exactly compensated by an inflated Type 1 error rate. The observed ratios by De Winter & Dodou (2015) show the ratio is the smallest for *p*-values between 0.001-0.009, and substantially higher for *p*-values between 0.011 and 0.049, with a relatively small increase in these 4 bins. This pattern can be reproduced just based on inflated Type 1 errors, but the required increase in Type 1 error rates over the 5 bins is very unlikely to occur when *p*-hacking.

The higher the average power of statistical tests, the more frequently small *p*-values will be observed if there is a true effect. This means there are more *p*-values between 0.021-0.029 than between 0.041-0.049 whenever the power is larger than 0. Without *p*-hacking, the number of Type 1 errors in each bin (e.g., between 0.001 and 0.009) should be 0.8% (it is 1% between 0 and 0.01). If we assume there were no inflated Type 1 error rates in 1990 (which is a conservative, albeit unlikely, estimate), the Type 1 error rates need to be increased to higher levels to reproduce the observed ratios, after selecting the average power of the studies, and the ratio of studies where H0 is true and H1 is true. It becomes extremely difficult to reconstruct both the observed absolute numbers and ratios.

One attempt to model the ratios (but not the absolute values) is presented in Table 4. The ratio of studies where H0 is true to studies where H1 is true is set to 1, and the average power is assumed to be 57.5%. The Type 1 error rate inflation over time is substantial, and the difference in the increase over the bins is not very typical, with a practically equal increase between 0.021-0.049. To achieve the ratios observed by De Winter & Dodou (2015) for comparisons between 2013 and years after 1990 the Type 1 error rate even needs to be inflated more strongly for *p*-values between 0.021-0.029 than for *p*-values between 0.041-0.049. Such a pattern of Type 1 error rate inflation is practically difficult to achieve, because questionable research practices (such as performing multiple analyses on the same data with different outlier criteria) produce a *p*-value distribution where higher *p*-values (e.g., 0.049) are observed more frequently than smaller *p*-values (e.g., 0.029). Thus, although it is not impossible to achieve the observed ratios purely by *p*-hacking (although it is very challenging to reconstruct both ratio's and absolute numbers), the required Type 1 error rate inflation over the 5 bins of *p*-values is unlikely to occur in real life for the observed ratios in most of the years. Furthermore, the required *average* inflation of Type 1 error rates across science needs to be substantial, close to three times the nominal Type 1 error rate. Alternatively, the Type 1 error rate inflation can be smaller, but one has to change the ratio of true effects to false effects that researchers examine by assuming researchers are substantially (e.g., five times) more likely to examine an idea that is false than an idea that is true. Neither scenario seems to be plausible.

Table 4. Absolute number of reconstructed Type 1 errors between 0.001-0.049 from 1990 to 2013.

	1990 true effects	2013 true effects	Type 1 error rate 1990	1990 Type 1 error	Type 1 error rate 2013	2013 Type 1 error	Reconstructed 1990/2013 Ratio
p0.001-p0.009	1814	47784	0.008	90	0.015	4449	6.66
p0.011-p0.019	492	12959	0.008	90	0.020	5932	7.89
p0.021-p0.029	319	8399	0.008	90	0.025	7415	9.40
p0.031-p0.039	238	6260	0.008	90	0.025	7415	10.14
p0.041-p0.049	189	4988	0.008	90	0.027	8008	11.30

To summarize, we can easily reconstruct the observed ratios by assuming a relatively small decrease in power over the years (e.g., from 55% to 42%). On the other hand, while increases in Type 1 error rates can be used to reconstruct the observed ratios, the pattern of inflated Type 1 errors across the 5 bins of *p*-values is unlikely to emerge in real life. Therefore, I conclude it is not very likely to be true that there is a 'surge of *p*-values between 0.041-0.049', nor that these data suggest there is an increase in questionable research practices over the last 25 years. The search for evidence of an increase in questionable research practices in science in general, or at least across a large number of studies, is starting to mirror the search for the ether. After repeatedly claiming to observe a rise in *p*-values just below 0.05 without providing substantial evidence for such a rise (De Winter & Dodou, 2015; Leggett et al., 2013; Masicampo & Lalande, 2012), it is time researchers investigating inflated Type 1 errors use better models, make better predictions, and collect better data.

I do not aim to suggest that the average decrease in power over time that I used in the reconstruction of the observed ratios reflects the true decrease in power over time. Other researchers are free to disagree with the specific parameters used to reconstruct the observed ratios. However, the current approach is an improvement over previous attempts to interpret differences in *p*-value distributions because it is based on a detailed model. Any criticisms on the suggestion that changes in power over time are a more likely explanation of the observed ratios than inflated Type 1 error rates should propose a better model before the current model can be abandoned. The proposed explanation for the observed *p*-value distribution contains clearly testable predictions, such as the leniency of reviewers and editors to accept marginally significant *p*-values as support for a hypothesis, and the

303 prediction than on average, power has decreased from 1990-2013.Testing these predictions in future
304 studies allow the current model to be either falsified or corroborated.

305 Analyzing huge numbers of *p*-values, which come from studies with large heterogeneity, will
306 not be able to provide any indication of the prevalence of questionable research practices, not even
307 when changes of *p*-value distributions are analyzed over time. A better approach seems to be to perform
308 targeted analyses of small sets of homogeneous studies, which might be able to yield support for *p*-
309 hacking (e.g., Lakens, 2014b). But more than anything else, the analyses in the present article point to
310 the fact that low statistical power and publication bias, and not *p*-hacking, are the biggest problems in
311 the scientific literature. Although it is important to control Type 1 error rates (e.g., Lakens, 2014c), it is
312 more important to design well-powered studies with high informational value (e.g., Lakens & Evers,
313 2014) and to reduce publication bias, for example by performing pre-registered studies that are
314 published regardless of the significance level of the results (e.g., Nosek & Lakens, 2014).

315 **Footnotes**

316 [1] The authors also analyze *p*-values with 2 digits (e.g., $p = 0.04$), which reveal similar patterns, but
317 here I focus on the three digit data, which focuses on *p*-values between for example 0.041-0.049
318 because trailing zeroes (e.g., $p = 0.040$) are rarely reported).

319 [2] These ratios should be 1, assuming all other parameters that determine *p*-value distributions remain
320 equal over time. As will be discussed later, this is not likely to be true, because there is reason to
321 assume a reduction in the average power over time. However, differences in power should also have a
322 greater impact on *p*-values below 0.05 than above 0.05.

323

References

van Assen, M. A., van Aert, R., & Wicherts, J. M. (in press). Meta-Analysis using effect size distributions of only statistically significant studies. *Psychological Methods*.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376.

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

Lakens, D. (2014a). What *p*-hacking really looks like: A comment on Masicampo and Lalande (2012). *The Quarterly Journal of Experimental Psychology*, (ahead-of-print), 1-4.

Lakens, D. (2014b). Professors are not elderly: Evaluating the evidential value of two social priming effects through *p*-curve analyses. *Available at SSRN*: http://ssrn.com/abstract=2381936

Lakens, D. (2014c). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*, 701-710. DOI: 10.1002/ejsp.2023.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of *p* values just below. 05. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2271-2279.

Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, *45(3)*, 137-141.

Lakens, D., & Evers, E. R. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, *9(3)*, 278-292.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105(2)*, 309-316.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve and effect size: correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666-681.

de Winter, J. C., & Dodou, D. (2015). A surge of *p*-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ*, *3*, e733.