

A peer-reviewed version of this preprint was published in PeerJ on 11 April 2016.

[View the peer-reviewed version](https://peerj.com/articles/1935) (peerj.com/articles/1935), which is the preferred citable publication unless you specifically need to cite this preprint.

Hartgerink CHJ, van Aert RCM, Nuijten MB, Wicherts JM, van Assen MALM. 2016. Distributions of p -values smaller than .05 in psychology: what is going on? PeerJ 4:e1935 <https://doi.org/10.7717/peerj.1935>

Distributions of p -values smaller than .05 in Psychology: What is going on?

Chris HJ Hartgerink, Robbie CM van Aert, Michèle B Nuijten, Jelte M. Wicherts, Marcel ALM van Assen

Previous studies provided mixed findings on peculiarities in p -value distributions in psychology. This paper examined 258,050 test results across 30,710 articles from eight high impact journals to investigate the existence of a peculiar prevalence of p -values just below .05 in the psychological literature, and a potential increase thereof over time. We indeed found evidence for a bump just below .05 in the distribution of exactly reported p -values in the journals *Developmental Psychology*, *Journal of Applied Psychology*, and *Journal of Personality and Social Psychology*, but the bump did not increase over the years and disappeared when using recalculated p -values. We found clear and direct evidence for the QRP "incorrect rounding of p -value" (John et al., 2012) in all psychology journals. Finally, we also investigated monotonic excess of p -values, an effect of certain QRPs that has been neglected in previous research, and developed two measures to detect this by modeling the distributions of statistically significant p -values. Using simulations and applying the two measures to the retrieved test results, we argue that, although one of the measures suggests the use of QRPs in psychology, it is difficult to draw general conclusions concerning QRPs based on modeling of p -value distributions.

Distributions of p -values smaller than .05 in Psychology: What is going on?

Chris H.J. Hartgerink¹, Robbie C.M. van Aert¹, Michèle B. Nuijten¹, Jelte M. Wicherts¹, and Marcel A.L.M. van Assen^{1,2}

¹Tilburg University, Warandelaan 2, 5037AB Tilburg

²Utrecht University, Padualaan 14, 3584 CH Utrecht

ABSTRACT

Previous studies provided mixed findings on peculiarities in p -value distributions in psychology. This paper examined 258,050 test results across 30,710 articles from eight high impact journals to investigate the existence of a peculiar prevalence of p -values just below .05 in the psychological literature, and a potential increase thereof over time. We indeed found evidence for a bump just below .05 in the distribution of exactly reported p -values in the journals *Developmental Psychology*, *Journal of Applied Psychology*, and *Journal of Personality and Social Psychology*, but the bump did not increase over the years and disappeared when using recalculated p -values. We found clear and direct evidence for the QRP "incorrect rounding of p -value" (John et al., 2012) in all psychology journals. Finally, we also investigated monotonic excess of p -values, an effect of certain QRPs that has been neglected in previous research, and developed two measures to detect this by modeling the distributions of statistically significant p -values. Using simulations and applying the two measures to the retrieved test results, we argue that, although one of the measures suggests the use of QRPs in psychology, it is difficult to draw general conclusions concerning QRPs based on modeling of p -value distributions.

Keywords: p -values, NHST, QRP, data peeking, Caliper test

INTRODUCTION

A set of p -values can be informative of the underlying effects that are investigated, but can also be indicative of potential research biases or questionable research practices (QRPs). Masicampo and Lalande (2012) found a bump of p -values just below .05 in three main psychology journals (i.e., *Journal of Personality and Social Psychology*, JPSP; *Journal of Experimental Psychology: General*, JEPG; *Psychological Science*, PS), which could be explained by research biases. A bump has occurred when p -values just below .05 are more prevalent than smaller p -values. The observation of a bump was one of several signals of a crisis of confidence in research findings in psychological science (Pashler and Wagenmakers, 2012; Ferguson, 2015). Leggett et al. (2013) later corroborated this bump of p -values for JPSP and JEPG, and observed that it was larger in 2005 than in 1965. Considering that research biases can lead to overemphasis on statistical significance, this result suggested that the state of psychology may have even deteriorated over the years. Additional corroboration in samples of published articles from various fields was provided by Head et al. (2015), who documented the bump of p -values below .05 in 1,048,575 articles across 16 disciplines including psychology. Ginsel et al. (2015) found similar biased reporting of p -values in medical abstracts, but noted the variety of potential causes (e.g., publication bias, fraud, selective reporting).

At the same time, other studies failed to find a bump of p -values below .05 (Jager and Leek, 2014; Krawczyk, 2015; Vermeulen et al., 2015). Reanalysis of data from Masicampo and Lalande (2012) and Head et al. (2015) indicated that the original results may have been confounded by publication bias and tendencies to round p -values (Lakens, 2015b; Hartgerink, 2015). Publication bias refers to the fact that the probability of getting published is higher for statistically significant results than for statistically nonsignificant results (Gerber et al., 2010; Franco et al., 2014). Publication bias only changes the p -value distribution above .05 and cannot cause a bump. Krawczyk (2015) analyzed a sample of around 5,000 psychology articles and found no bump in p -values that were *recalculated* on the basis of reported test statistics and degrees of freedom (cf. Bakker and Wicherts, 2011). However, he did observe a bump

35 for reported p -values. As such, this highlights an important difference between reported p -values and
36 recalculated p -values, and stresses the need to distinguish both types of results when studying signs of
37 questionable research practices.

38 In this paper we differentiate between two forms of peculiar prevalence of p -values just below .05; a
39 bump and monotonic excess. Monotonic excess signifies a higher than expected frequency of p -values just
40 below .05, but in the absence of a bump, as in Figure 1b below.

41 In light of the aforementioned conflicting findings and interpretations, the present paper attempts to
42 answer two questions: (1) Does a bump or monotonic excess of p -values below .05 exist in psychology?
43 and (2) Did evidence for a bump increase over time in psychology? We chose to focus on psychology
44 because of the availability of an extensive database on statistical results in psychology (used in Nuijten
45 et al., 2015) and because discussions on research practices are particularly salient in this discipline (e.g.,
46 Pashler and Wagenmakers, 2012; John et al., 2012; Simmons et al., 2011; Wagenmakers et al., 2012;
47 Asendorpf et al., 2013).

48 First we clarify how the two research questions relate to questionable research practices (QRPs). QRPs
49 are defined as practices that are detrimental to the research process (Panel on Scientific Responsibility
50 and the Conduct of Research, 1992), with a recent focus on those which "increase the likelihood of
51 finding support for a false hypothesis" (p.524 John et al., 2012). Several QRPs related to significance
52 testing are known to affect p -values of statistical tests and consequently the decisions based on these
53 tests. Specifically, particular QRPs may yield results that are just significant and can create a bump
54 of p -values, such as ad hoc exclusion of outliers (Bakker and Wicherts, 2014), repeatedly sampling
55 new participants and checking the results (i.e., data peeking, Armitage et al., 1969), including various
56 combinations of covariates until a significant result is reached, or operationalizing a measure in different
57 ways until significance is reached (Simmons et al., 2011).

58 These QRPs have been used by many researchers at least once in their career. For instance, data
59 peeking and the ad hoc exclusion of outliers were admitted by 63% and 38% of psychological researchers,
60 respectively (John et al., 2012). On the other hand, other QRPs mainly yield very small and (clearly)
61 significant p -values, such as analyzing multiple conditions or correlated variables and selecting only the
62 smallest p -value out of this set of analyses (van Aert et al., 2015; Ulrich and Miller, 2015) and do not lead
63 to a bump. To summarize, different QRPs may differently affect the distribution of statistically significant
64 p -values.

65 In the absence of QRPs, the distribution of significant p -values can be expected to have a certain shape.
66 Under the null-hypothesis all p -values are equally probable (i.e., follow a uniform distribution). If there
67 is truly an effect, smaller p -values are more likely than larger p -values (i.e., the distribution decreases
68 monotonically in the p -value). Consequently, because some hypotheses are false and some are true, the
69 distribution of observed p -values arises from a mixture of uniform and right-skewed distributions and
70 should also decrease monotonically.¹ Deviation from a monotonically decreasing distribution (i.e., a
71 bump) could indicate evidence of QRPs that aim to obtain just significant results. Hence answering our
72 research questions of whether a bump exists and whether this bump changed over time may also inform
73 us on the prevalence of these particular QRPs and its development over time.

74 However, there are at least two problems with using p -value distributions to examine the prevalence
75 of QRPs. First, as we previously argued, not all QRPs lead to a bump of p -values just below .05. Hence,
76 examining the distribution of p -values just below .05 will not inform us on the prevalence of QRPs that do
77 not aim to obtain just significant results but yield mainly small and clearly significant p -values (van Aert
78 et al., 2015; Ulrich and Miller, 2015). Second, the QRPs yielding just significant results do not necessarily
79 result in a non-monotonic p -value distribution, that is, a distribution with a *bump*. For instance, consider
80 Figure 1 that shows the result of simulations done for data peeking, which is known to result in mainly
81 just significant p -values (Armitage et al., 1969; Lakens, 2015b; Wagenmakers, 2007). The dashed lines in
82 both panels correspond to 20 million simulated p -values under the null-hypothesis and a medium effect
83 size ($d = .5$), respectively, in a two-sample t -test with 24 participants per group. The solid lines show
84 the distributions of 20 million simulated p -values under these same effect sizes and designs, but after a
85 maximum of three rounds of data peeking with each round adding 1/3 of the original sample size. Figure 1
86 illustrates that data peeking may result in non-monotonic excess (i.e., bump; left panel), but can also cause

¹One exception to this rule is when the alternative hypothesis is wrongly specified, that is, if the true effect size is negative whereas the alternative hypothesis states that the true effect is positive. In this case the distribution of the p -value is left-skewed and monotonically increasing.

87 *monotonic excess* (right panel), even if all researchers use data peeking. Specifically, if all underlying
 88 effects are genuinely and substantially different from zero (right panel), data peeking will generally not
 89 lead to a bump below .05. In the present paper, we therefore examine the peculiar prevalence of p -values
 90 just below .05 by both investigating the presence of a bump and monotonic excess in distributions of
 91 statistically significant results.

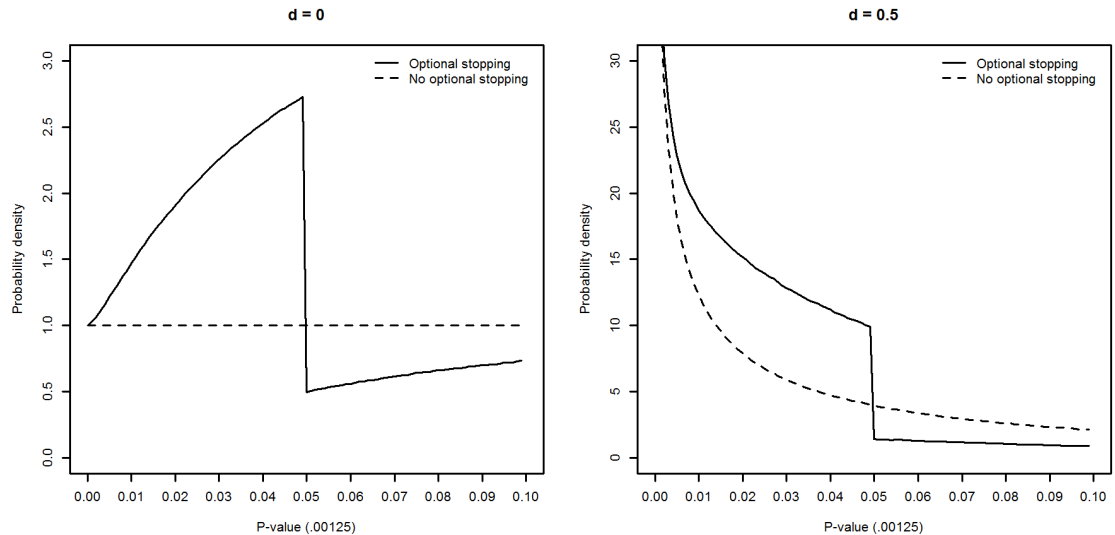


Figure 1. Distributions of 20 million p -values each, when $d = 0$ (bump; left) and $d = .5$ (monotonic excess; right), given data peeking (solid) or no data peeking (dashed). Simulations were run for two-sample t -tests with $n_k = 24$. For data peeking, a maximum of three rounds of additional sampling occurred if the result was nonsignificant, with each round adding $1/3$ of the original sample size.

92 In answering our research questions, whether a bump or monotonic excess exist and whether the bump
 93 changed over time, we improve previous studies on four dimensions. First, we eliminate the distortive
 94 effects of publication bias on the p -value distribution by inspecting only statistically significant results.
 95 Second, we use a large dataset on p -values from entire articles instead of only p -values from abstracts
 96 (as in Jager and Leek, 2014; de Winter and Dodou, 2015). Third, we distinguish between reported and
 97 recalculated p -value distributions for the same set of test results and show that this distinction affects
 98 answers to the two questions because of common mismatches (Bakker and Wicherts, 2011). Fourth, we
 99 fit analytic models to p -value distributions to investigate for monotonic excess, where previous research
 100 only investigated whether there was non-monotone excess (i.e., a bump).

101 Publication bias distorts the p -value distribution, but distortions caused by this bias should not be
 102 confounded with distortions caused by QRPs. Publication bias refers to the selective publication of
 103 disproportionate amounts of statistically significant outcomes (Gerber et al., 2010; Franco et al., 2014).
 104 Publication bias contributes to a higher frequency of p -values just below .05 relative to the frequency
 105 of p -values just above .05, but only does so by decreasing the frequency of p -values larger than .05.
 106 Masicampo and Lalande (2012) and de Winter and Dodou (2015) indeed found this relatively higher
 107 frequency, which is more readily explained by publication bias, which affects the distribution of p -values
 108 larger than .05. QRPs that lead to a bump affect only the distribution of p -values smaller than .05 (Lakens,
 109 2015b). We focus only on the distribution of significant p -values, because this distribution is affected by
 110 QRPs that cause a bump or monotonic excess and not, or to a much lesser extent, by publication bias.

111 The second extension is the use of more extensive data for psychology than previously used to inspect
 112 QRPs that cause a bump or monotonic excess, improving our ability to examine the prevalence of QRPs.
 113 Masicampo and Lalande (2012) and Leggett et al. (2013) manually collected p -values from a relatively
 114 small set of full research articles (i.e., 3,627 and 3,701), whereas Jager and Leek (2014) and de Winter
 115 and Dodou (2015) used automated extraction of p -values from only the abstracts of research papers.
 116 However, p -values from abstracts are not representative for the population of p -values from the entire
 117 paper (Benjamini and Hechtlinger, 2014; Ioannidis, 2014), even though some have argued against this

118 (Pautasso, 2010). Our large scale inspection of full-text articles is similar to papers by Head et al. (2015)
 119 and Krawczyk (2015), but with the addition of being able to recalculate all p -values from the test statistics
 120 extracted from the text.

121 Third, we examine the prevalence of QRPs that cause a bump or monotonic excess by investigating
 122 both reported and the accompanying recalculated p -values. Previous studies do not fully compare results
 123 from reported p -values and recalculated p -values. This distinction is relevant, because reported p -values
 124 are subject to reporting bias such as rounding errors, particularly relevant around the .05 threshold. Such
 125 reporting biases result in inaccurate p -value distributions. For example, there is evidence that reporting
 126 errors that affect statistical significance occur in approximately 10-15% of papers in psychology (i.e.,
 127 gross inconsistencies Bakker and Wicherts, 2011; García-Berthou and Alcaraz, 2004; Nuijten et al., 2015;
 128 Veldkamp et al., 2014). The advantage of analyzing recalculated p -values is that they contain more
 129 decimals than typically reported and correct reporting errors. Some previous studies analyzed reported
 130 p -values (de Winter and Dodou, 2015; Jager and Leek, 2014; Head et al., 2015), whereas others looked
 131 at recalculated p -values (Masicampo and Lalande, 2012) or a mix of reported and recalculated (Leggett
 132 et al., 2013). Only Krawczyk (2015) used both reported and recalculated p -values for a subset of the
 133 data, and they found that the peculiar prevalence below .05 disappeared when the recalculated data were
 134 used. Hence, this distinction between reported and recalculated p -values allows us to distinguish between
 135 peculiarities due to reporting errors and peculiarities due to QRPs such as data peeking.

136 Fourth, we examine the prevalence of p -values just below .05 by taking into account various models
 137 to test and explain characteristics of p -value distributions. We applied tests and fitted models to p -values
 138 below .05, in two ways. We first applied the non-parametric Caliper test (Gerber et al., 2010) comparing
 139 frequencies of p -values in an interval just below .05 to the frequency in the adjacent lower interval; a
 140 higher frequency in the interval closest to .05 is evidence for QRPs that seek to obtain just significant
 141 results. The Caliper test has also been applied to examine publication bias, by comparing just significant
 142 to just nonsignificant p -values (Kühberger et al., 2014), and to detect QRPs (Head et al., 2015). However,
 143 the Caliper test can only detect a bump but not monotonic excess, as illustrated by the distributions of
 144 p -values in Figure 1. Therefore, we also attempted to model the distribution of significant p -values in
 145 order to investigate for all forms of excess (i.e., both a bump and monotonic excess), and illustrate the
 146 results and difficulties of this approach.

147 In short, this paper studies the distribution of significant p -values in four ways. First, we verified
 148 whether a bump is present in *reported* p -values just below .05 with the Caliper test. Second, to examine
 149 how reporting errors might influence p -value distributions around .05, we analyzed only the recalculated
 150 p -values corresponding to those reported as .05. Third, we used the Caliper test to examine if a bump
 151 effect is present in *recalculated* p -values and whether evidence for a bump changed over time. Finally,
 152 we modeled the distribution of significant recalculated p -values in an attempt to also detect a monotonic
 153 excess of p -values below .05.

154 DATA AND METHODS

155 Data

156 We investigated the p -value distribution of research papers in eight high impact psychology journals (also
 157 used in Nuijten et al., 2015). These eight journals were selected due to their high-impact across different
 158 subfields in psychology and their availability within the Tilburg University subscriptions. This selection
 159 also encompasses the journals covered by Masicampo and Lalande (2012) and Leggett et al. (2013). A
 160 summary of the downloaded articles is included in Table 1.

Journal	Acronym	Timespan	Articles downloaded	Articles with extracted results (%)	APA results extracted
Developmental Psychology	DP	1985-2013	3,381	2,607 (77%)	37,658
Frontiers in Psychology	FP	2010-2013	2,126	702 (33%)	10,149
Journal of Applied Psychology	JAP	1985-2013	2,782	1,638 (59%)	15,134
Journal of Consulting and Clinical Psychology	JCCP	1985-2013	3,519	2,413 (69%)	27,429
Journal of Experimental Psychology General	JEPG	1985-2013	1,184	821 (69%)	18,921
Journal of Personality and Social Psychology	JPSP	1985-2013	5,108	4,346 (85%)	101,621
Public Library of Science	PLOS	2000-2013	10,303	2,487 (24%)	31,539
Psychological Science	PS	2003-2013	2,307	1,681 (73%)	15,654
		<i>Total</i>	30,710	16,695 (54%)	258,105

Table 1. Articles downloaded, articles with extracted APA results, and number of extracted APA test results per journal.

161 For these journals, our sample included articles published from 1985 through 2013 that were available
 162 in HTML format. For the PLOS journals, HTML versions of articles were downloaded automatically
 163 with the `rplos` package (v0.3.8; Chamberlain et al., 2015). This package allows an R user to search
 164 the PLOS database as one would search for an article on the website.² We used this package to retrieve
 165 search results that include the subject ‘psychology’ for (part of) an article. For all other journals, HTML
 166 versions of articles were downloaded manually by the first author.

167 APA test results were extracted from the downloaded articles with the R package `statcheck` (v1.0.1;
 168 Epskamp and Nuijten, 2015). The only requirement for this package to operate is a supply of HTML (or
 169 PDF) files of the articles that are to be scanned and `statcheck` extracts all test results reported according
 170 to the standards of the American Psychological Association (APA; American Psychological Association,
 171 2010). This format is defined as test results reported in the following order: the test statistic and degrees
 172 of freedom (encapsulated in parentheses) followed by the p -value (e.g., $t(85) = 2.86, p = .005$). This
 173 style has been prescribed by the APA since at least 1983 (American Psychological Association, 1983,
 174 2001), with the only relevant revision being the precision of the reported p -value, changing from two
 175 decimal places to three decimal places in the sixth edition from 2010. `statcheck` extracts t , F , χ^2 , Z
 176 and r results reported in APA style. Additional details on the validity of the `statcheck` package can be
 177 found in Nuijten et al. (2015).

178 From the 30,710 downloaded papers, `statcheck` extracted 258,105 test results. We removed 55
 179 results, because these were impossible test results (i.e., $F(0,55) = \dots$ or $r > 1$). The final dataset thus
 180 included 258,050 test results. The extracted test results can have four different formats, where test results
 181 or p -values are reported either exactly (e.g., $p = .042$) or inexactly (e.g., $p < .05$). Table 2 shows the
 182 composition of the dataset, when split across these (in)exactly reported p -values and (in)exactly reported
 183 test results.

	Exact test statistic	Inexact test statistic	
Exact p -value	68,776	274	69,050 (27%)
Inexact p -value	187,617	1,383	189,000 (73%)
	256,393 (99.36%)	1,657 (0.64%)	258,050 (100%)

Table 2. Composition of extracted APA test results with respect to exact and inexact reporting of p -values or test statistics.

184 From this dataset, we selected six subsets throughout our analyses to investigate our research questions
 185 regarding a bump below .05. We analyzed (i) all reported p -values ($N = 258,050$) for a bump in their
 186 distribution just below .05. Subsequently we analyzed (ii) only exactly reported p -values ($N = 69,050$).
 187 It is possible that reporting or rounding errors have occurred among the reported p -values. To investigate
 188 the degree to which this happens at $p = .05$, we analyzed (iii) exactly reported test statistics that are
 189 accompanied by an exactly reported p -value of .05 (i.e., $p = .05$). This subset contains 2,470 results. To
 190 debilitate rounding errors and other factors influencing the reporting of p -values (e.g., Ridley et al., 2007),
 191 we also investigated the recalculated p -value distribution with (iv) p -values that were accompanied by
 192 exactly reported test statistics ($N = 256,393$). To investigate whether evidence for a bump differs for
 193 inexactly and exactly reported p -values, (v) 68,776 exactly reported test statistics with exactly reported
 194 p -values were analyzed. Finally, we used (vi) all recalculated p -values in 0-.05 for t , r , and $F(df_1 = 1)$
 195 values to model the effect size distribution underlying these p -values to investigate evidence of both a
 196 bump and monotonic excess.

197 Methods

198 We used the Caliper test and two new measures to examine if the observed p -value distribution shows
 199 evidence for a bump or monotonic excess below .05. We applied the two measures to the observed
 200 p -value distribution and we examined their performance to detect a bump or monotonic excess using a
 201 simulation study on data peeking. Data peeking was chosen because it is one of the most frequently used
 202 and well-known QRPs. Below, we explain the Caliper test, how the p -value distributions are modeled
 203 with the two new measures, and describe the design of the simulation study in more detail.

²We note there are minor differences in the number of search results from the PLOS webpage and the `rplos` package for equal searches. This is due to differences in the default search database for the webpage and the package. For technical details on this issue, see <https://github.com/ropensci/rplos/issues/75>

204 **Caliper test**

205 In order to test for a bump of p -values just below .05, we applied the Caliper test (e.g., Gerber et al., 2010;
206 Kühberger et al., 2014). This proportion test compares the frequencies of p -values in two intervals, such
207 as the intervals .04-.045 and .045-.05. Let Pr denote the proportion of p -values of the interval .045-.05.
208 Then, independent of the population effect sizes underlying the p -values, Pr should not be higher than .5
209 in any situation because the p -value distribution should be monotone decreasing. Hence $Pr > .5$ signifies
210 a bump of p -values just below .05.

211 We carried out one-tailed binomial proportion tests, with $H_0 : Pr \leq .5$ and $H_1 : Pr > .5$. For example,
212 if 40 and 60 p -values are observed in the intervals .04-.045 and .045-.05, respectively, then $Pr = .6$ and
213 the binomial test results in p -value = .0284, suggesting evidence for a bump below .05. We applied the
214 Caliper test to the reported p -values (subsets one through three as described in the previous section) and
215 recalculated p -values (subsets four and five), both for the entire dataset and each of the eight psychology
216 journals.

217 The Caliper test requires specifying the width of the intervals that are to be compared. For reported
218 p -values, which are frequently rounded to two-decimal values, we selected the intervals (.03875-.04] and
219 [.04875-.05) because there is a strong preference to report p -values to the second decimal in research
220 papers (see also Hartgerink, 2015). For recalculated p -values we used the same interval width as used by
221 Masicampo and Lalande (2012) and Leggett et al. (2013), which is .00125, corresponding to a comparison
222 of intervals (.0475-.04875] and [.04875-.05). Note that rounding is not a problem for recalculated p -values.
223 Considering that some journals might show small frequencies of p -values in these intervals, we also
224 carried out Caliper tests with interval widths of .0025, .005, and .01. Note that, on the one hand, increasing
225 interval width increases the statistical power of the Caliper test because more p -values are included in
226 the test, but on the other hand also decreases power because Pr is negatively related to interval width
227 whenever p -values correspond to tests of non-zero population effects. In other words, a bump just below
228 .05 will tend more and more towards a monotonically decreasing distribution as the binwidth increases.

229 To verify if evidence for a bump of p -values increased over time, we fitted a linear trend to proportion
230 Pr of the Caliper test with binwidths .00125, .0025, .005, and .01. We computed these proportions for
231 each year separately, for both the total dataset and per journal. Time was centered at the start of data
232 collection, which was 1985 except for PLOS (2000), PS (2006; due to 0 p -values in the considered
233 interval for preceding years), and FP (2010). The value .5 was subtracted from all Pr values, such that the
234 intercept of the trend corresponds to the bump of p -values at the start of data collection, where 0 means
235 no bump. A positive linear trend signifies an increase in the bump of p -values below .05 over time.

236 **Measures based on p -value distributions**

237 Figure 1 demonstrates that data peeking has a different effect on the p -value distribution depending on
238 the true effect size. The distribution after data peeking does not monotonically increase for $d = 0$ (left
239 panel), whereas it does increase monotonically for $d = 0.5$ (right panel). Consequently, the Caliper test
240 will signal a bump of p -values for $d = 0$ (i.e., it will detect a bump), but not for $d = 0.5$.

241 We examined how we may be able to detect both a bump and monotonic excess of p -values below
242 .05. Figure 1 indicates that, for p -values close to zero (e.g., $\leq .00125$) the p -value distributions with
243 data peeking (solid lines) closely match the p -value distributions without data peeking (dashed lines). In
244 other words, data-peeking in studies with initially nonsignificant p -values rarely results in tiny significant
245 p -values, but more often in p -values larger than .00125. The basic idea of this analysis is therefore to
246 estimate the ‘true’ effect size distribution using only these tiny p -values (i.e., $\leq .00125$), assuming that
247 none or a very small proportion of these p -values were affected by by data-peeking.

248 We examined the performance of two measures to detect a bump or monotonic excess of p -values
249 below .05. The first method compares the effect sizes estimated on p -values smaller than .00125 to effect
250 sizes estimated using all p -values smaller than .05. The idea of this first method is that increasing the
251 frequency of just-significant p -values *decreases* the effect size estimate. Indeed, the more right-skewed
252 the p -value distribution, the higher the effect size estimate when keeping constant studies’ sample sizes
253 (Simonsohn et al., 2014; van Assen et al., 2015). According to the first method, there is evidence
254 suggestive of data peeking (or other QRPs leading to a bump of p -values just below .05) if the effect size
255 estimate is considerably lower when based on all p -values than when based on only p -values $\leq .00125$.

The second method yields a measure of excess of p -values just below .05, for either a bump or
monotonic excess, by comparing the observed frequency of p -values in the interval .00125-.05 to the

predicted frequency of p -values in that interval. This prediction is based on the effect size estimated using the p -values smaller than .00125. If the ratio of observed over expected p -values is larger than 1, referred to as statistic D , then this could indicate data peeking. Statistic D is calculated as

$$D = \frac{P_{.00125}^o}{1 - P_{.00125}^o} \times \frac{1 - P_{.00125}^e}{P_{.00125}^e} \quad (1)$$

256 with $P_{.00125}^o$ and $P_{.00125}^e$ representing the proportion of p -values lower than .00125 observed and expected,
257 respectively. Note that D is an odds ratio.

258 Modeling p -value distributions

For both measures from the previous section the expected p -value distribution needs to be derived and compared to the observed p -value distribution. The observed p -value distribution of the psychology data is based on all exactly reported statistics with test statistics t , r , and $F(1, df_2)$, because these readily provide the same effect measure. We used the Fisher transformed correlation, ρ_F , as effect size measure. The distribution of the Fisher transformed correlation is approximated well by the normal distribution, with Fisher transformation

$$\rho_F = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) \quad (2)$$

and standard error $\frac{1}{\sqrt{N-3}}$ or $\frac{1}{\sqrt{df_2-1}}$. $F(1, df_2)$ and t values can be transformed to correlations using

$$r = \frac{\frac{F \times df_1}{df_2}}{\frac{F \times df_1}{df_2} + 1} \quad (3)$$

259 where $F = t^2$.

The expected p -value distribution was estimated under the assumption that the true effect size, Fisher transformed correlation ρ_F , is normally distributed with μ_{ρ_F} and standard deviation τ_{ρ_F} . The two parameters were estimated by minimizing the χ^2 -statistic

$$\chi_{j-1}^2 = \sum_{j=1}^J \frac{(rf_j^o - rf_j^e)^2}{rf_j^e} \quad (4)$$

with rf_j^o and rf_j^e being the relative frequency of observed and expected p -values in interval j , respectively. Minimization was done with the `optim()` function in R, where $\hat{\tau}$ was truncated to be positive. Interval j is defined as $(I_{j-1}, I_j) = ((j-1)x, jx)$, with width $x = .00025$ whenever only the significant p -values lower than .00125 were modeled (resulting in 5 intervals); .00125 when modeling all significant p -values (i.e., $p \leq .05$, 40 intervals); .025 when modeling all p -values (also 40 intervals). The relative frequencies are conditional probabilities. For instance, rf_2^o is the proportion of observed p -values in interval $(I_1 = .00025, I_2 = .0005)$ whenever p -values lower than .00125 are examined. Expected relative frequency rf_j^e is computed as

$$rf_j^e = \frac{\sum_{k=1}^K P(I_{j-1} \leq p_k \leq I_j | N_k; \hat{\rho}_F; \hat{\tau}_{\rho_F})}{\sum_{k=1}^K P(p_k \leq I_j | N_k; \hat{\rho}_F; \hat{\tau}_{\rho_F})} \quad (5)$$

260 with the summation over all K significant test statistics. P corresponds to the probability that a p -value of
261 study k (i.e., p_k) is in a certain interval, which depends on the study sample size N_k and the two estimated
262 parameters of the effect size distribution (i.e., $\hat{\rho}_F$, $\hat{\tau}_{\rho_F}$).

263 Design of simulation study

264 To examine the potential of the two measures to detect data peeking, their performance was examined
265 on simulated data with and without data peeking. We used a two-group between-subjects design with
266 24 participants per group ($n_k = 24$), and compared their means using a t -test. The performance of both
267 measures was examined as a function of true effect size μ (0; 0.2; 0.5; 0.8) and heterogeneity τ (0; 0.15).

268 In the data peeking conditions, data were simulated as follows: means and variances per group were
 269 simulated and a two-sample t -test was conducted. If this t -test was statistically significant (i.e., $p \leq .05$),
 270 the p -value was stored, otherwise the data peeking procedure was started. In this data peeking procedure,
 271 one-third of the original sample size was added to the data before conducting another two-sample t -test.
 272 This data peeking procedure was repeated until a statistically significant result was obtained or three
 273 rounds of additive sampling had taken place (see osf.io/x5z6u for simulation functions). The simulations
 274 were stopped if 1,000,000 studies with a p -value below .1 were obtained for each combination of μ and τ .

275 RESULTS AND DISCUSSION

276 In this section, we report the results of our analyses in the following order for the subsets: all reported
 277 p -values (258,050 results), exactly reported p -values (69,050 results), p -values erroneously reported as
 278 equal to .05 (2,470 results), all recalculated p -values based on exactly reported test statistics (256,393
 279 results), recalculated p -values based on exactly reported test statistics and exactly reported p -values
 280 (68,776 results), and the modeling of p -value distributions based on recalculated p -values 0-.00125 and
 281 0-.05 (54,561 results and 127,509, respectively). These analyses apply the Caliper test to investigate
 282 evidence of a possible bump below .05. Subsequently, the results of the two measures are presented based
 283 on all recalculated p -values.

284 Reported p-values

285 Figure 2 shows the distribution for all reported p -values (i.e., 258,050; white bars) and exactly reported p -
 286 values (i.e., 69,050; blue bars). Results of the Caliper test indicate (i) there is a bump just below .05 when
 287 considering all reported p -values in bins .03875-.04 versus .04875-.05, $N = 45,667$, $Pr = 0.905$, $p < .001$
 288 and (ii) there is still a bump, but less so, when considering only exactly reported p -values in these bins,
 289 $N = 4,900$, $Pr = 0.547$, $p < .001$. The difference in bumps between these two subsets can be explained
 290 by the amount of p -values that are reported as $< .05$, which is 86% of all p -values reported as exactly
 291 equal to .05 and 14% of all reported p -values.

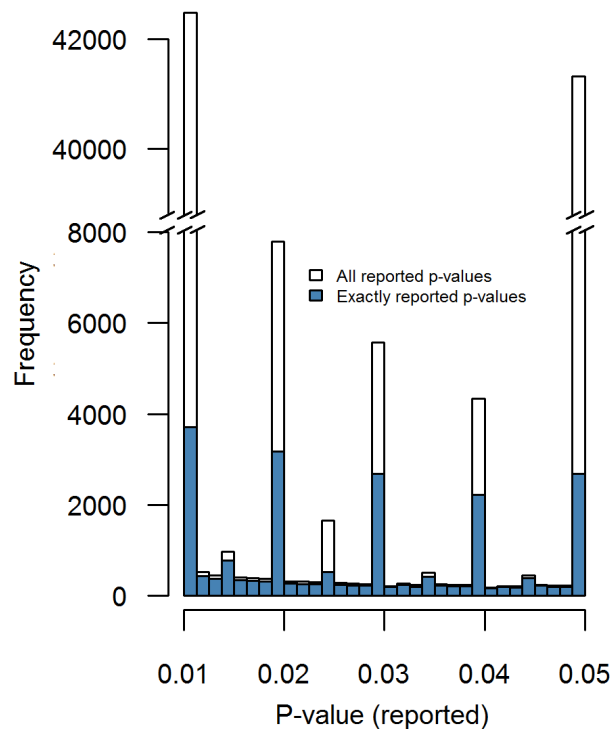


Figure 2. Distributions of all reported p -values (white) and exactly reported p -values (blue) across eight psychology journals. Binwidth = .00125.

292 To investigate whether this observed bump below .05 across exactly reported p -values originates from

293 one or multiple journals, we performed the Caliper test on the exactly reported p -values per journal. Table
 294 3 shows the results for these tests. The results indicate that there is sufficient and reliable evidence for
 295 a bump below .05 (i.e., $Pr > .5$) for the journals DP and JPSP and sufficient evidence, but debatable
 296 reliability for JAP, where the results depend on the binwidth. However, the other five journals show no
 297 evidence for a bump below .05 in exactly reported p -values at all. In other words, the bump below .05 in
 298 exactly reported p -values is mainly driven by the journals DP, JAP, and JPSP.

Binwidth	0.00125				0.0025				0.005				0.01			
	x	N	Pr	p	x	N	Pr	p	x	N	Pr	p	x	N	Pr	p
All	2,682	4,900	0.547	< .001	2,881	5,309	0.543	< .001	3,308	6,178	0.535	< .001	4,218	8,129	0.519	< .001
DP	319	531	0.601	< .001	336	567	0.593	< .001	383	653	0.587	< .001	464	843	0.55	0.002
FP	96	193	0.497	0.557	105	227	0.463	0.884	141	304	0.464	0.906	215	458	0.469	0.912
JAP	78	131	0.595	0.018	82	137	0.599	0.013	85	154	0.552	0.113	101	183	0.552	0.092
JCCP	246	517	0.476	0.874	267	562	0.475	0.889	308	641	0.48	0.848	395	823	0.48	0.882
JEPG	147	285	0.516	0.318	159	310	0.513	0.346	195	375	0.52	0.235	258	509	0.507	0.395
JPSP	1,252	2,097	0.597	< .001	1,310	2,207	0.594	< .001	1,408	2,399	0.587	< .001	1,623	2,869	0.566	< .001
PLOS	307	649	0.473	0.921	366	760	0.482	0.854	489	1,000	0.489	0.766	744	1,558	0.478	0.964
PS	237	497	0.477	0.859	256	539	0.475	0.886	299	652	0.459	0.984	418	886	0.472	0.957

Table 3. Caliper test for exactly reported p -values per journal for different binwidths. x = frequency of p -values in .05 minus binwidth through .05, N = total frequency of p -values across both intervals in the comparison, $Pr = x/N$, $p = p$ -value of the binomial test. Significant results ($\alpha = .05$, one-tailed) indicating excess of p -values just below .05 and are reported in bold.

299 The Caliper test results for reported p -values indicate two things: (i) in exactly reported p -values
 300 severely distort the p -value distribution, and (ii) a bump below .05 is also found when only considering
 301 exactly reported p -values. Because inexact reporting of p -values causes excess at certain points of the
 302 p -value (e.g., the significance threshold .05; Ridley et al., 2007), we recommend only inspecting exactly
 303 reported p -values when examining p -value distributions.

304 Considering only exactly reported p -values, there is sufficient evidence for a bump below .05 in the
 305 journals DP, JAP, and JPSP, but not in the remaining five journals (i.e., FP, JCCP, JEPG, PLOS, PS). A
 306 tentative explanation of the bump of p -values just below .05 for DP, JAP, and JPSP may be that QRPs that
 307 aim to obtain barely significant results are more frequent in the fields of these journals. However, another
 308 explanation may be that scientists in these fields are more prone to exactly report p -values just below .05
 309 (e.g., to emphasize they are really smaller than .05) than p -values considerably smaller than .05.

310 Recalculated p -value distributions

311 Recalculated when reported $p = .05$

312 Results for reported p -values remain inconclusive with regard to the distribution of p -values, due to
 313 potential rounding or errors (Bakker and Wicherts, 2011; Nuijten et al., 2015; Veldkamp et al., 2014).
 314 Rounding and errors could result in an over-representation of p -values $\leq .05$. To investigate the plausibility
 315 of this notion, we inspected recalculated p -values when $p = .05$ was reported. Figure 3 indicates that
 316 p -values that were reported as .05 show remarkable spread when recalculated, which indicates that the
 317 reported p -value might frequently be rounded or incorrect, assuming that the reported test statistics are
 318 correct. More specifically, 67.45% of p -values reported as .05 were larger than .05 when recalculated and
 319 32.55% were smaller than .05. This percentage does not greatly vary across journals (range 58.8%-73.4%
 320 compared to 67.45%). Taking into account rounding possibilities (i.e., widening the range of correct
 321 p -values to .045-.055), these percentages become 13.81% and 7.85%, respectively, meaning that at least
 322 21.66% of the p -values reported as .05 was incorrectly reported. In comparison, p -values reported as
 323 $p = .04$, $p = .03$, or $p = .02$ show smaller proportions of downward rounding when compared to $p = .05$
 324 (i.e., 53.33%, 54.32%, 50.38%, respectively compared to 67.45%). When taking into account potential
 325 rounding errors in the initial reporting of p -values, the discrepancy remains but to a smaller extent (i.e.,
 326 11.74%, 9.57%, 8.03%, respectively compared to 13.81%). These results provide direct evidence for
 327 the QRP "incorrect rounding of p -value" (John et al., 2012), which contributes to a bump or monotonic
 328 excess just below .05.

329 The discrepancy between recalculated p -values and p -values reported as equal to .05 highlights
 330 the importance of using recalculated p -values when underlying effect distributions are estimated as in
 331 p -uniform and p -curve (van Assen et al., 2015; Simonsohn et al., 2014). When interested in inspecting the
 332 p -value distribution, reported p -values can substantially distort the p -value distribution, such that results

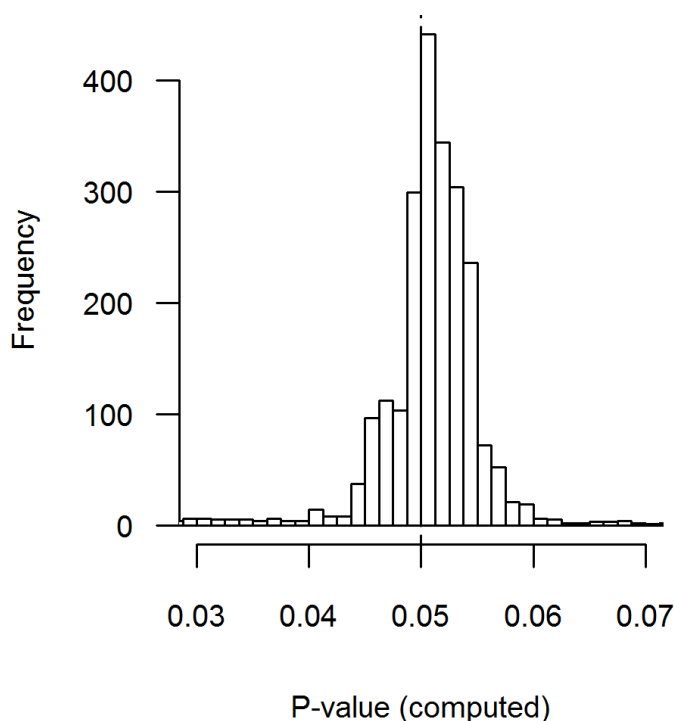


Figure 3. Distribution of recalculated p -values where the p -value is reported as $p = .05$. 9.7% of the results fall outside the range of the plot, with 3.6% at the left tail and 6.1% at the right tail. Binwidth = .00125

333 become biased if we rely solely on the reported p -value. Such a discrepancy indicates potential rounding
 334 of p -values, erroneous reporting of p -values, or strategic reporting of p -values. The p -value distortions
 335 can be (partially) corrected for by recalculating p -values based on reported test statistics. Additionally,
 336 potential distortions to the distribution at the third decimal place due to the rounding of p -values to the
 337 second decimal (Hartgerink, 2015) is also solved by recalculating p -values. We continue with recalculated
 338 p -values in our following analyses.

339 **Recalculated p -values**

340 Figure 4 shows the distribution of all recalculated p -values (i.e., set of 256,393 results) and of recalculated
 341 p -values whenever the reported p -value is exact (i.e., set of 68,776 results). The recalculated p -value
 342 distribution is markedly smoother than the reported p -value distribution (see Figure 2) due to the absence
 343 of rounded p -values.

344 After inspecting all recalculated p -values, we did not observe a bump just below .05, $N = 2,808$, $Pr =$
 345 $.5$, $p = 0.508$. When we analyzed the recalculated p -values per journal (Table 4), there is no evidence
 346 for a bump below .05 in any of the journals. Additionally, we inspected all recalculated p -values that
 347 resulted from exactly reported p -values. For this subset we did observe a bump below .05, $N = 809$, $Pr =$
 348 0.564 , $p = 0.000165$ (blue histogram in Figure 4) for the smallest binwidth (i.e., .00125), but this effect
 349 was not robust across larger binwidths, as shown in Table 5. This table also specifies the results for a
 350 bump below .05 per journal, with sufficient evidence of a bump only in JPSP. This finding, however, was
 351 only observed for binwidths .00125 and .0025, not for larger binwidths. Considering the results from
 352 the recalculated p -values, there is sparse evidence for the presence of a bump below .05, opposed to
 353 widespread evidence (Masicampo and Lalande, 2012; Leggett et al., 2013; Head et al., 2015). Moreover,
 354 interpretation of the bump for JPSP is not straightforward; it may also be that authors of JPSP are more
 355 prone to report exact test statistics if the p -value is just below .05 than whenever p -values are considerably
 356 smaller than .05.

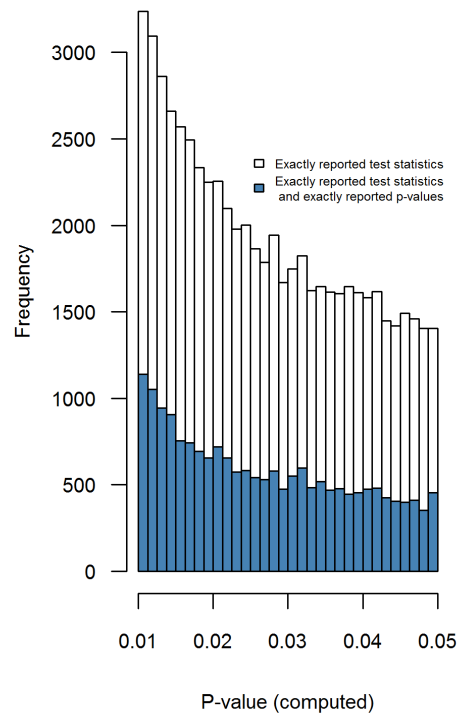


Figure 4. Recalculated p -values for exactly reported test statistics (white bars), and recalculated p -values for exactly reported test statistics where p -values are also exactly reported (blue bars). Binwidth = .00125

357 Excessive significance over time

358 The regression results of the development of a bump below .05 over time, based on recalculated p -values,
 359 are shown in Table 6. Results indicate that there is no evidence for a linear relation between time in years
 360 and the degree to which a bump of p -values below .05 is present across the different binwidths (only
 361 results for binwidth .00125 are presented; results for the other binwidths available at <http://osf.io/96kbc/>).
 362 Conversely, for PLOS there is some evidence for a minor increase of a bump throughout the years
 363 ($b = .072, p = .039$), but this result is not robust for binwidths .0025, .005, and .01. These results contrast
 364 with Leggett et al. (2013), who found a linear relation between time and the degree to which a bump
 365 occurred for JEPG and JPSP. Hence, our findings contend the increase of a bump below .05 for the period
 366 1965-2005 in psychology (Leggett et al., 2013). In other words, our results of the Caliper test indicate
 367 that, generally speaking, there is no evidence for an increasing prevalence of p -values just below .05 or of
 368 QRPs causing such a bump in psychology.

Binwidth	0.00125				0.0025				0.005				0.01			
	x	N	Pr	p	x	N	Pr	p	x	N	Pr	p	x	N	Pr	p
All	1,404	2,808	0.5	0.508	2,808	5,761	0.487	0.973	5,761	11,824	0.487	0.997	11,824	25,142	0.47	> .999
DP	184	382	0.482	0.779	382	829	0.461	0.989	829	1,710	0.485	0.9	1,710	3,579	0.478	0.996
FP	30	69	0.435	0.886	69	172	0.401	0.996	172	376	0.457	0.956	376	799	0.471	0.955
JAP	73	145	0.503	0.5	145	270	0.537	0.124	270	556	0.486	0.765	556	1,168	0.476	0.952
JCCP	160	308	0.519	0.265	308	633	0.487	0.763	633	1,267	0.5	0.522	1,267	2,706	0.468	> .999
JEPG	81	164	0.494	0.593	164	332	0.494	0.608	332	683	0.486	0.778	683	1,535	0.445	> .999
JPSP	640	1,268	0.505	0.379	1,268	2,557	0.496	0.668	2,557	5,174	0.494	0.802	5,174	10,976	0.471	> .999
PLOS	125	260	0.481	0.752	260	541	0.481	0.828	541	1,170	0.462	0.995	1,170	2,544	0.46	> .999
PS	111	212	0.524	0.268	212	427	0.496	0.577	427	888	0.481	0.88	888	1,835	0.484	0.919

Table 4. Caliper test for exactly recalculated p -values per journal for different binwidths. x = frequency of p -values in .05 minus binwidth through .05, N = total frequency of p -values across both intervals in the comparison, $Pr = x/N$, p = p -value of the binomial test. Significant results ($\alpha = .05$, one-tailed) indicating excess of p -values just below .05 and are reported in bold.

Binwidth	0.00125				0.0025				0.005				0.01			
	<i>x</i>	<i>N</i>	<i>Pr</i>	<i>p</i>	<i>x</i>	<i>N</i>	<i>Pr</i>	<i>p</i>	<i>x</i>	<i>N</i>	<i>Pr</i>	<i>p</i>	<i>x</i>	<i>N</i>	<i>Pr</i>	<i>p</i>
All	456	809	0.564	<.001	809	1,617	0.5	0.5	1,617	3,403	0.475	0.998	3,403	7,402	0.46	1
DP	46	87	0.529	0.334	87	185	0.47	0.811	185	358	0.517	0.281	358	756	0.474	0.932
FP	15	27	0.556	0.351	27	87	0.31	>.999	87	192	0.453	0.915	192	437	0.439	0.995
JAP	8	20	0.4	0.868	20	29	0.69	0.031	29	65	0.446	0.839	65	141	0.461	0.844
JCCP	43	78	0.551	0.214	78	161	0.484	0.682	161	364	0.442	0.988	364	780	0.467	0.971
JEPG	27	50	0.54	0.336	50	98	0.51	0.46	98	209	0.469	0.834	209	479	0.436	0.998
JPSP	184	305	0.603	<.001	305	547	0.558	0.004	547	1,117	0.49	0.764	1,117	2,451	0.456	>.999
PLOS	76	149	0.51	0.435	149	323	0.461	0.926	323	698	0.463	0.978	698	1,470	0.475	0.975
PS	57	93	0.613	0.019	93	187	0.497	0.558	187	400	0.468	0.912	400	888	0.45	0.999

Table 5. Caliper tests for exactly recalculated and exactly reported *p*-values per journal, including alternative binwidths. *x* = frequency of *p*-values in .05 minus binwidth through .05, *N* = total frequency of *p*-values across both intervals in the comparison, *Pr* = *x*/*N*, *p* = *p*-value of the binomial test. Significant results ($\alpha = .05$, one-tailed) indicating excess of *p*-values just below .05 and are reported in bold.

	Timespan	Coefficient	Estimate	SE	<i>t</i>	<i>p</i>
All	1985-2013	Intercept	0.007	0.017	0.392	0.698
All		Years (centered)	-0.001	0.001	-0.492	0.627
DP	1985-2013	Intercept	-0.043	0.056	-0.769	0.448
DP		Years (centered)	0.001	0.003	0.193	0.849
FP	2010-2013	Intercept	-0.182	0.148	-1.233	0.343
FP		Years (centered)	0.055	0.079	0.694	0.560
JAP	1985-2013	Intercept	0.041	0.081	0.504	0.619
JAP		Years (centered)	-0.001	0.005	-0.208	0.837
JCCP	1985-2013	Intercept	0.077	0.058	1.315	0.200
JCCP		Years (centered)	-0.006	0.004	-1.546	0.134
JEPG	1985-2013	Intercept	-0.022	0.124	-0.176	0.862
JEPG		Years (centered)	0.001	0.007	0.097	0.924
JPSP	1985-2013	Intercept	-0.002	0.027	-0.062	0.951
JPSP		Years (centered)	0.000	0.002	-0.005	0.996
PLOS	2006-2013	Intercept	-0.382	0.114	-3.344	0.016
PLOS		Years (centered)	0.072	0.027	2.632	0.039
PS	2003-2013	Intercept	0.081	0.078	1.045	0.323
PS		Years (centered)	-0.009	0.013	-0.669	0.520

Table 6. Linear regression coefficients as a test of increasing excess of *p*-values just below .05. Intercept indicates the degree of excess for the first year of the estimated timespan ($> 0 =$ excess). Significant results ($\alpha = .05$, two-tailed) are reported in bold.

369 **Results of two measures based on modeling p -value distributions**

370 ***Data of eight psychology journals***

371 Figure 5 depicts the observed p -value distribution and the expected p -value distribution corresponding to
 372 the fitted effect size distribution based on p -values $\leq .00125$. Estimates for p -values $\leq .05$ were effect
 373 size $\hat{\rho}_F = 0$ and heterogeneity $\hat{\tau}_{\rho_F} = .183$, and $\hat{\rho}_F = .149$ and $\hat{\tau}_{\rho_F} = .106$ for p -values $\leq .00125$. Misfit
 374 between observed and expected p -value distribution for $p \leq .00125$ was minor ($\chi^2 = 4.1$), indicating that
 375 the observed p -values $\leq .00125$ were well approximated by the estimated effect size distribution.

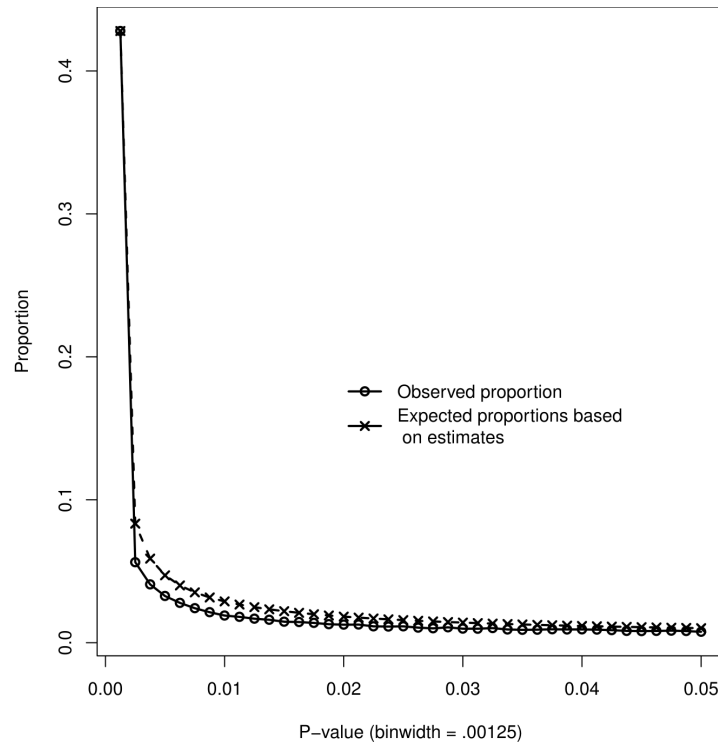


Figure 5. Observed proportions of p -values (circles) and expected proportions of p -values based on $\hat{\rho}_F$ and $\hat{\tau}_{\rho_F}$ estimated from 0-.00125 (crosses).

376 Our first measure suggests practices leading to a monotonic excess of p -values below .05, because
 377 the estimated effect size based on all significant p -values (i.e., 0) is much smaller than the supposedly
 378 unbiased estimate based on only the very small p -values (i.e., .183). Moreover, assuming that effect
 379 sizes are normally distributed with $\rho_F = 0$ and $\tau_{\rho_F} = .183$, combined with the degrees of freedom of the
 380 observed effects, implies that only 27.5% of all effects would be statistically significant. However, of all
 381 reported p -values, 74.7% were statistically significant, but this difference may at least partly be caused by
 382 other factors such as publication bias. It is highly unlikely that the average true effect size underlying
 383 statistically significant results in psychology is truly zero. It remains undecided, however, whether this
 384 very low estimate is mainly due to QRPs leading to a downward bias of the effect size estimate, or to a
 385 misspecification of the model, an issue we revisit later in the paper.

386 For the second measure that compares the ratio of observed and expected p -values below .05, we
 387 found $D = .701$, which does not suggest data peeking but *under-reporting* of p -values (29.9%) in the
 388 p -value interval .00125-.05. The simulation results that follow below, however, demonstrated that the
 389 measure D performs badly under effect size heterogeneity. Since heterogeneity is underlying the observed
 390 data, we conclude that the measure D is not useful for investigating evidence of a bump or monotonic
 391 excess of p -values.

392 ***Simulation study***

393 Table 7 shows the results of the two measures for data simulated with and without data peeking. The
 394 column headers show the mean (i.e., μ) and heterogeneity (i.e., τ) of the simulated conditions, with the

395 corresponding ρ_F and τ_{ρ_F} on the Fisher transformed correlation scale. The first set of rows shows the
 396 results for the data simulated without data peeking, of which we discuss the results first.

p-values		$\tau = 0$				$\tau = .15$					
		$\mu = 0$ $\rho_F = 0$	$\mu = .2$ $\rho_F = .099$	$\mu = .5$ $\rho_F = .247$	$\mu = .8$ $\rho_F = .390$	$\mu = 0$ $\rho_F = 0$	$\mu = .2$ $\rho_F = .099$	$\mu = .5$ $\rho_F = .247$	$\mu = .8$ $\rho_F = .390$		
Without data peeking	0-1	$\hat{\rho}_F$	0	0.103	0.258	0.413	0	0.103	0.258	0.413	
		$\hat{\tau}_{\rho_F}$	0	0	0	0	0.077	0.077	0.077	0.077	
		Misfit χ^2	0	0	0	0	0	0	0	0	
	0-.05	$\hat{\rho}_F$	0	0.103	0.258	0.413	0	0.103	0.258	0.413	
		$\hat{\tau}_{\rho_F}$	0	0	0	0.001	0.077	0.077	0.077	0.077	
		Misfit χ^2	0	0	0	0	0	0	0	0	
	0-.00125	$\hat{\rho}_F$	0	0.103	0.258	0.413	0.1	0.107	0.259	0.413	
		$\hat{\tau}_{\rho_F}$	0	0	0	0.001	0.025	0.076	0.077	0.077	
		Misfit χ^2	0	0	0	0	0	0	0	0	
		D	1	1	1	1	1.205	1.006	1.003	1.001	
	With data peeking	0-.05	$\hat{\rho}_F$	0	0	0.117	0.345	0	0	0.075	0.360
			$\hat{\tau}_{\rho_F}$	0	0	0	0.038	0	0.055	0.137	0.091
Misfit χ^2			126,267.4	50,298.4	696.6	101.6	14,867.6	1,209.5	576.3	340.6	
		N	759,812	811,296	936,517	994,974	434,660	525,023	707,650	889,681	
0-.00125		$\hat{\rho}_F$	0	0.075	0.218	0.366	0.066	0.161	0.283	0.402	
		$\hat{\tau}_{\rho_F}$	0	0	0	0	0.036	0	0	0.012	
		Misfit χ^2	6.9	3.2	7.1	11.8	2	1.9	2.6	2.1	
		N	9,729	21,576	95,615	350,482	14,791	34,530	124,991	366,875	
		D	1.977	1.976	1.835	1.166	1.628	1.620	1.472	1.164	

Table 7. Results of parameter estimation of the distribution of effect sizes and measures of data peeking as a function of population effect size (μ, ρ_F), population heterogeneity (τ), and data peeking, for the simulated data. Results are based on all p -values 0-1, p -values $\leq .05$, and $\leq .00125$. $\hat{\rho}_F$ = estimated population effect, $\hat{\tau}_{\rho_F}$ = estimated population heterogeneity, misfit 0-.05 = misfit of estimates based on p -values 0-.05, misfit 0-.00125 = misfit of estimates based on p -values 0-.00125 (bold indicates $p < .05$), N = number of results included in estimation, D = comparison of observed- and expected p -value frequencies.

397 The results for the data without data peeking inform us on (i) whether the effect size distribution
 398 parameters can accurately be recovered using only very small ($\leq .00125$) or small p -values ($\leq .05$), and
 399 (ii) if both measures accurately signal no data peeking. Note that ρ_F is slightly overestimated due to
 400 categorizing the p -value distribution into 40 categories: the estimates based on all p -values (i.e., $\hat{\rho}_F$, first
 401 row) are slightly larger than the population parameter (i.e., ρ_F , column headers).

402 Answering the first question of accurate parameter estimates, whenever there is no heterogeneity
 403 (i.e., $\tau_{\rho_F} = 0$) both ρ_F and τ_{ρ_F} are accurately recovered. When heterogeneity is non-zero, the parameters
 404 were also accurately recovered, but not when $\rho_F = 0$. Here, ρ_F was overestimated (equal to .1) and τ_{ρ_F}
 405 underestimated (.025 rather than the true .077), while at the same time the misfit was negligible.

406 The latter result, that the effect is overestimated under heterogeneity when $\rho_F = 0$, is explained
 407 by the fact that a p -value distribution can accurately be modeled with an infinite range of negatively
 408 correlated values of ρ_F and τ_{ρ_F} . An increase in ρ_F yields a more right-skewed distribution, which is
 409 hardly distinguishable from the right-skewed distribution caused by an increase in τ_{ρ_F} . The similar effects
 410 of both parameters on the fitted p -value distribution already hint at potential problems for both measures,
 411 because performance of these measures is dependent on accurate estimates of these parameters.

412 With respect to the second question, whether the measures accurately signal the absence of data
 413 peeking, the first measure does so in both homo- and heterogeneous conditions, whereas the second
 414 measure correctly signals absence only under homogeneity. The first measure signals data peeking if the
 415 estimate of ρ_F is smaller when based on $p \leq .05$ than on $p \leq .00125$. Previously, we already noted that
 416 effect size estimates were identical to population effect sizes under homogeneity, and equal or *larger*
 417 when based on $p \leq .00125$ under heterogeneity. This suggests that the first measure behaves well if there
 418 is no data peeking (but see the conclusion section). The second measure, D , performed well (i.e., was
 419 equal to 1) under homogeneity, but incorrectly suggested data peeking under heterogeneity. For instance,
 420 $D = 1.205$ for $\rho_F = 0$ and $\tau = .15$, which suggests that 20.5% more p -values were observed in the interval
 421 .00125-.05 than were expected based on the $\hat{\rho}_F$ estimate even though no data peeking occurred. The
 422 explanation for the breakdown of the performance of D is that the parameters of the effect size distribution
 423 were not accurately recovered, overestimating the average effect size and underestimating heterogeneity
 424 based on small p -values. This yields a lower expected frequency of higher p -values (between .00125 and

425 .05), thereby falsely suggesting data peeking.

426 The last rows present the results obtained when data peeking does occur. First consider the estimates
427 of ρ_F and the performance of the first measure of data peeking. The estimates of ρ_F confirm that data
428 peeking results in underestimation, particularly if the average true effect size is not large (i.e., $\mu = .2$ or
429 $.5$). Moreover, downward bias of ρ_F decreases when it is estimated on p -values $\leq .00125$ than on $\leq .05$,
430 accurately signaling data peeking with the first measure. For instance, if $\rho_F = .099$ and $\tau = 0$, $\hat{\rho}_F = .075$
431 when based on p -values $\leq .00125$ and $\hat{\rho}_F = 0$ when based on p -values $\leq .05$. Together with the good
432 performance of this measure under no data peeking, these results suggest that the first measure may be
433 useful to detect data keeping in practice.

434 Consider the estimates of τ_{ρ_F} and the performance of D . Similar to conditions under no data peeking,
435 heterogeneity is grossly underestimated when using p -values $\leq .00125$. Hence D cannot be expected
436 to perform well under data peeking. Although D -values seem to correctly signal data peeking in all
437 conditions and decrease as expected when the effect size increases, these values do not correspond to
438 the actual values of data peeking. For instance, consider the condition with $\mu = .5$ and $\tau_{\rho_F} = .15$; of the
439 582,659 simulated p -values in interval $.00125$ -.05, 106,241 p -values were obtained through data-peeking,
440 which yields a true $D = 1.223$, which is very different from the estimated $D = 1.472$ in Table 7.

441 Finally, consider the (mis)fit of the estimated p -value distribution. Despite the considerable downward
442 bias in heterogeneity estimate $\hat{\tau}_{\rho_F}$, the simulated p -value distribution is mostly well approximated by
443 the expected p -value distribution, as indicated by the small values of the χ^2 statistic for p -values in
444 0-.00125. Hence, good fit again does not imply accurate parameter estimates. The misfit of the estimated
445 distribution for p -values $\leq .05$ is indicated by large χ^2 -values, particularly when the p -value distribution
446 is not monotonically decreasing (which is the case for, e.g., $\mu = 0$).

447 To conclude, this simulation study showed that under true homogeneity both measures of data peeking
448 can accurately signal both absence and presence of data peeking. However, under true heterogeneity,
449 heterogeneity is underestimated and the performance of D breaks down, while results suggest that
450 comparing estimates of average effect size, the first measure, may still accurately signal both the absence
451 and presence of data peeking.

452 LIMITATIONS AND CONCLUSION

453 Before concluding, some limitations of our method to collect p -values need to be addressed. First,
454 `statcheck` (Epskamp and Nuijten, 2015; Nuijten et al., 2015), the R package used to collect the
455 observed data, extracts all APA test results reported in the text of an article, but not those reported in
456 tables. Hence, our selection of results is potentially not representative of all reported results, but this
457 most likely does not affect results. Second, our analysis assumed that test statistics other than p -values
458 were accurately reported. If test statistics and degrees of freedom are incorrectly reported, recalculated
459 p -values are wrong as well. We identified some erroneous test statistics (e.g., $df_1 = 0$ and $r > 1$), but
460 do not know how often these errors occur and how they may have affected our results. We assumed that
461 p -value errors were made due to the overemphasis on them in current day research.

462 In light of conflicting findings and interpretations, we aimed to provide final answers to the questions
463 (1) Does a bump or monotonic excess of p -values below $.05$ exist in psychology? and (2) Did evidence
464 for a bump increase over time in psychology? Answering these research questions may inform us on
465 the prevalence of QRPs and its development over time in psychology. Using `statcheck`, we extracted
466 and analyzed 258,050 test results conforming to APA-style across 30,710 articles from eight high impact
467 journals in psychology, and distinguished between results with inexactly reported p -values, exactly
468 reported p -values, and recalculated p -values. The basic idea underlying our analyses is that QRPs distort
469 the p -value distribution. We argued that only some QRPs yield an excess of p -values just below $.05$, and
470 show that QRPs sometimes yield a bump and sometimes only monotonic excess of p -values just below
471 $.05$. We used the Caliper test to test for a bump, and suggested two measures to examine monotonic excess.

472 Starting with the existence of a bump in psychology, we drew the following conclusions. First,
473 *inexactly* reported p -values are not useful for analyses of p -value distributions. Second, a bump in *exactly*
474 reported p -values indeed exists in psychology journals DP, JAP, and JPSP. QRPs leading to just significant
475 p -values can explain these bumps, but we also cannot rule out the explanation that scientists in these
476 particular journals are more prone to exactly report p -values just below $.05$ (e.g., to emphasize they are
477 really smaller than $.05$) than p -values considerably smaller than $.05$. Third, contradicting Leggett et al.
478 (2013), the bump and evidence of a bump in psychology did not increase over the years. Fourth, when

479 analyzing only the *exactly* reported p -values equal to .05, clear and direct evidence was obtained for the
480 QRP "incorrect rounding of p -value" (John et al., 2012). Evidence of this QRP, which contributed to
481 the bump in exactly reported p -values in psychology, was found in all psychology journals. Fifth, after
482 removing reporting errors and analyzing the *recalculated* reported p -values, evidence of a bump was
483 found only for JPSP. Again, this may have been caused by QRPs or by scientists being more prone to
484 report all test statistics when p -values are just below .05 than if they are considerably smaller than zero.

485 The conclusions obtained with the two measures investigating monotonic and non-monotonic excess are
486 not satisfactory. First, performance of both measures is dependent on accurately recovering parameters
487 of the effect size distribution, which turned out to be difficult; estimates of effect size heterogeneity
488 and average effect size are highly correlated and unstable when based on only statistically significant
489 findings. Second, simulations show that one of the measures, D , does not accurately assess the QRP data
490 peeking when effect sizes are heterogeneous. Third, even though performance of the second measure
491 (i.e., difference between effect sizes based on contaminated and supposedly uncontaminated p -values)
492 is affected by estimation problems, it correctly signaled data peeking in the simulations. Fourth, when
493 applying the second measure to the observed distribution of significant p -values in psychology, the
494 measure found evidence of monotonic excess of p -values; the average effect size estimate based on all
495 these p -values was 0, which seems very unrealistic, and suggests the use of QRPs in psychology leading
496 to p -values just below .05.

497 Notwithstanding the outcome of the second measure, suggesting QRPs that cause monotonic excess,
498 we do not consider it as direct evidence of such QRPs in psychology. Lakens (p.3; 2015) suggests that "it
499 is essential to use a model of p -value distributions before drawing conclusions about the underlying reasons
500 for specific distributions of p -values extracted from the scientific literature." We explicitly modeled the
501 effect size distribution and by using the degrees of freedom of test results also model the effect sizes'
502 power and the p -value distribution. But we fear this is not and cannot be sufficient. First of all, we could
503 not accurately recover the effect size distribution under heterogeneity in our simulation study, even if all
504 assumptions of our model were met. This rendered measure D unfruitful when there is heterogeneity,
505 and severely limits the usefulness of the second measure that compares estimated average effect sizes.
506 Second, devising other models may yield other results and thereby other interpretations (Benjamini and
507 Hechtlinger, 2014; Goodman, 2014; Lakens, 2015a; de Winter and Dodou, 2015).

508 Results of all the aforementioned models are most likely not robust to violations of their assumptions.
509 For instance, we assume a normal distribution of true effect sizes. This assumption is surely violated, since
510 the reported p -values arise from a mixture of many different types of effects, such as very large effects
511 (manipulation checks), effects corresponding to main hypotheses, and zero effects ('control' variables).
512 Additionally, consider the QRPs themselves; we examined the effect of only one QRP, data peeking, in
513 one of its limited variants. Other QRPs exist that also increase the prevalence of p -values just below .05,
514 such as multiple operationalizations of a measure and selecting the first one to be significant. Other QRPs
515 even increase the frequency of very small p -values (van Aert et al., 2015). We deem it impossible to
516 exhaustively model QRPs and their effects, considering the difficulties we show for a single QRP that is
517 clearly defined. To conclude, we fear that Gelman and O'Rourke (2014) may be right when suggesting
518 that drawing conclusions with regard to any QRP based on modeling p -value distributions obtained from
519 automatically extracted results is unfruitful.

520 On the other hand, we do recommend modeling effect size and p -value distributions of results
521 that all intend to test the same hypothesis, to prevent contamination by irrelevant test results (Bishop
522 and Thompson, 2015; Simonsohn et al., 2015). Examples of methods that focus on similar results are
523 p -uniform (van Assen et al., 2015) and p -curve (Simonsohn et al., 2014), which model statistically
524 significant statistics pertaining to one specific effect and estimate the effect size based on these statistics
525 while correcting for publication bias. Further research should reveal if both methods can also be used
526 to detect and correct for p -hacking in the context of estimating one particular effect size. Preliminary
527 results suggest, however, that detection and correcting for p -hacking based on statistics alone is rather
528 challenging (van Aert et al., 2015).

529 REFERENCES

530 American Psychological Association (1983). *Publication manual of the American Psychological Association*.
531 American Psychological Association, Washington, DC, 3rd edition.

- 532 American Psychological Association (2001). *Publication manual of the American psychological associa-*
533 *tion*. American Psychological Association, Washington, DC, 5th edition.
- 534 American Psychological Association (2010). *Publication manual of the American Psychological Associa-*
535 *tion*. American Psychological Association, Washington, DC, 6th edition.
- 536 Armitage, P., McPherson, C. K., and Rowe, B. C. (1969). Repeated significance tests on accumulating
537 data. *Journal of the Royal Statistical Society. Series A*, 132(2):235–244.
- 538 Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S.,
539 Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G.,
540 Weber, H., and Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology.
541 *European journal of personality*, 27(2):108–119.
- 542 Bakker, M. and Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals.
543 *Behavior research methods*, 43(3):666–678.
- 544 Bakker, M. and Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the type I error
545 rate in independent samples t tests: the power of alternatives and recommendations. *Psychological*
546 *methods*, 19(3):409–427.
- 547 Benjamini, Y. and Hechtlinger, Y. (2014). Discussion: An estimate of the science-wise false discovery
548 rate and applications to top medical journals by Jager and Leek. *Biostatistics*, 15(1):13–16.
- 549 Bishop, D. V. and Thompson, P. A. (2015). Problems in using text-mining and p-curve analysis to detect
550 rate of p-hacking. Technical Report e1550, PeerJ PrePrints.
- 551 Chamberlain, S., Boettiger, C., and Ram, K. (2015). rplos: Interface to the search 'API' for 'PLOS'
552 journals.
- 553 de Winter, J. C. and Dodou, D. (2015). A surge of p-values between 0.041 and 0.049 in recent decades
554 (but negative results are increasing rapidly too). *PeerJ*, 3:e733.
- 555 Epskamp, S. and Nuijten, M. (2015). statcheck: Extract statistics from articles and recompute p-values.
- 556 Ferguson, C. J. (2015). 'everybody knows psychology is not a real science': Public perceptions of
557 psychology and how we can improve our relationship with policymakers, the scientific community, and
558 the general public. *The American psychologist*, 70(6):527–542.
- 559 Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: Unlocking
560 the file drawer. *Science*, 345(6203):1502–1505.
- 561 García-Berthou, E. and Alcaraz, C. (2004). Incongruence between test statistics and P values in medical
562 papers. *BMC medical research methodology*, 4:13.
- 563 Gelman, A. and O'Rourke, K. (2014). Discussion: Difficulties in making inferences about scientific truth
564 from distributions of published p-values. *Biostatistics*, 15(1):18–23.
- 565 Gerber, A. S., Malhotra, N., Dowling, C. M., and Doherty, D. (2010). Publication bias in two political
566 behavior literatures. *American Politics Research*, 38(4):591–613.
- 567 Ginsel, B., Aggarwal, A., Xuan, W., and Harris, I. (2015). The distribution of probability values in
568 medical abstracts: an observational study. *BMC research notes*, 8(1):721.
- 569 Goodman, S. N. (2014). Discussion: An estimate of the science-wise false discovery rate and application
570 to the top medical literature. *Biostatistics*, 15(1):23–7.
- 571 Hartgerink, C. H. J. (2015). Reanalyzing Head et al. (2015): No widespread p-hacking after all?
- 572 Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The extent and
573 consequences of p-hacking in science. *PLoS biology*, 13(3):e1002106.
- 574 Ioannidis, J. P. A. (2014). Discussion: Why "an estimate of the science-wise false discovery rate and
575 application to the top medical literature" is false. *Biostatistics*, 15(1):28–36.
- 576 Jager, L. R. and Leek, J. T. (2014). An estimate of the science-wise false discovery rate and application to
577 the top medical literature. *Biostatistics*, 15(1):1–12.
- 578 John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research
579 practices with incentives for truth telling. *Psychological science*, 23(5):524–532.
- 580 Krawczyk, M. (2015). The search for significance: A few peculiarities in the distribution of p values in
581 experimental psychology literature. *PLoS one*, 10(6):e0127872.
- 582 Kühberger, A., Fritz, A., and Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on
583 the correlation between effect size and sample size. *PLoS one*, 9(9):e105825.
- 584 Lakens, D. (2015a). On the challenges of drawing conclusions from p-values just below 0.05. *PeerJ*,
585 3:e1142.
- 586 Lakens, D. (2015b). What p-hacking really looks like: A comment on Masicampo and LaLonde (2012).

- 587 *Quarterly journal of experimental psychology*, 68(4):829–832.
- 588 Leggett, N. C., Thomas, N. A., Loetscher, T., and Nicholls, M. E. R. (2013). The life of p: “just significant”
589 results are on the rise. *Quarterly journal of experimental psychology*, 66(12):2303–2309.
- 590 Masicampo, E. J. and Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly*
591 *journal of experimental psychology*, 65(11):2271–2279.
- 592 Nuijten, M. B., Hartgerink, C. H. J., Van Assen, M. A. L. M., Epskamp, S., and Wicherts, J. M. (2015).
593 The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*.
594 Panel on Scientific Responsibility and the Conduct of Research (1992). *Responsible science, volume I:*
595 *Ensuring the integrity of the research process*. National Academies Press (US), Washington, DC.
- 596 Pashler, H. and Wagenmakers, E.-J. (2012). Editors’ introduction to the special section on replicability in
597 psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6):528–530.
- 598 Pautasso, M. (2010). Worsening file-drawer problem in the abstracts of natural, medical and social science
599 databases. *Scientometrics*, 85(1):193–202.
- 600 Ridley, J., Kolm, N., Freckelton, R. P., and Gage, M. J. G. (2007). An unexpected influence of widely
601 used significance thresholds on the distribution of reported p-values. *Journal of evolutionary biology*,
602 20(3):1082–1089.
- 603 Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed
604 flexibility in data collection and analysis allows presenting anything as significant. *Psychological*
605 *science*, 22(11):1359–1366.
- 606 Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of*
607 *experimental psychology: General*, 143(2):534–547.
- 608 Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2015). Better p-curves. *Journal of Experimental*
609 *Psychology: General*, 144:1146–1152.
- 610 Ulrich, R. and Miller, J. (2015). p-hacking by post hoc selection with multiple opportunities: Detectability
611 by skewness test?: Comment on simonsohn, nelson, and simmons (2014). *Journal of experimental*
612 *psychology. General*, 144(6):1137–1145.
- 613 van Aert, R. C. M., Wicherts, J. M., and van Assen, M. A. L. M. (2015). Conducting meta-analyses
614 on p-values: Reservations and recommendations for applying p-uniform and p-curve. *Manuscript*
615 *submitted for publication*.
- 616 van Assen, M. A. L. M., van Aert, R. C. M., and Wicherts, J. M. (2015). Meta-analysis using effect size
617 distributions of only statistically significant studies. *Psychological methods*, 20:293–309.
- 618 Veldkamp, C. L. S., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A. L. M., and Wicherts, J. M.
619 (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science.
620 *PloS one*, 9(12):e114876.
- 621 Vermeulen, I., Beukeboom, C. J., Batenburg, A., Avramiea, A., Stoyanov, D., de Velde, B. v., and
622 Oegema, D. (2015). Blinded by the light: How a focus on statistical “significance” may cause p-value
623 misreporting and an excess of p-values just below .05 in communication science. *Communication*
624 *methods and measures*, 9(4):253–279.
- 625 Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic*
626 *bulletin & review*, 14(5):779–804.
- 627 Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., and Kievit, R. A. (2012). An
628 agenda for purely confirmatory research. *Perspectives on psychological science: a journal of the*
629 *Association for Psychological Science*, 7(6):632–638.