The Quarterly Journal of Experimental Psychology

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pqje20

A peculiar prevalence of p values just below .05

E. J. Masicampo [a] & Daniel R. Lalande [b]

[a] Department of Psychology, Wake Forest University, Winston-Salem, NC, USA

[b] Department of Health Sciences, Université du Québec à Chicoutimi, Chicoutimi, QC, Canada

Accepted author version posted online: 13 Jul 2012. Version of record first published: 02 Aug 2012

To cite this article: E. J. Masicampo & Daniel R. Lalande (2012): A peculiar prevalence of p values just below .05, The Quarterly Journal of Experimental Psychology, DOI:10.1080/17470218.2012.711335

To link to this article: http://dx.doi.org/10.1080/17470218.2012.711335

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY
2012, i*First*, 1–9

Psychology Press
Taylor & Francis Group

A peculiar prevalence of *p* values just below .05

E. J. Masicampo[1], and Daniel R. Lalande[2]

[1]Department of Psychology, Wake Forest University, Winston-Salem, NC, USA
[2]Department of Health Sciences, Université du Québec à Chicoutimi, Chicoutimi, QC, Canada

In null hypothesis significance testing (NHST), *p* values are judged relative to an arbitrary threshold for significance (.05). The present work examined whether that standard influences the distribution of *p* values reported in the psychology literature. We examined a large subset of papers from three highly regarded journals. Distributions of *p* were found to be similar across the different journals. Moreover, *p* values were much more common immediately below .05 than would be expected based on the number of *p* values occurring in other ranges. This prevalence of *p* values just below the arbitrary criterion for significance was observed in all three journals. We discuss potential sources of this pattern, including publication bias and researcher degrees of freedom.

Keywords: Statistics; Statistical inference; Hypothesis testing.

The psychology literature is meant to comprise scientific observations that further people's understanding of the human mind and human behaviour. However, due to strong incentives to publish, the main focus of psychological scientists may often shift from practising rigorous and informative science to meeting standards for publication. One such standard is obtaining statistically significant results. In line with null hypothesis significance testing (NHST), for an effect to be considered statistically significant, its corresponding *p* value must be less than .05. The present paper examined whether biases linked to meeting that standard are evident in the psychological findings that are ultimately published. We examined whether an unusually high number of published effects have corresponding *p* values just below .05, as evidence of biases linked to achieving statistical significance.

NHST and reliance on *p*

Null hypothesis significance testing (NHST) is the most frequently used data analysis method in psychology and many other scientific disciplines (Kline, 2004). The null hypothesis is a statement of no relationship between variables or no effect of some experimental manipulation or intervention. When using NHST, one computes the probability *p* of finding an effect as extreme as, or more extreme than, the finding *F* observed given that the null hypothesis H_0 is true, or $p(F|H_0)$ (Nickerson, 2000; Trafimow, 2003). If the analysis reveals a *p* value equal to or below the arbitrary criterion of .05, the effect is considered statistically significant. In other words, it would be highly improbable to obtain such results if the null hypothesis were true. The null hypothesis is therefore rejected in

Correspondence should be addressed to E. J. Masicampo, E. J. Masicampo, 415 Greene Hall, P.O. Box 7778 Reynolda Station, Department of Psychology, Wake Forest University, Winston-Salem, NC 27109, USA or to Daniel R. Lalande, 6846 Clark St., Montreal, PQ, H2S 3E9. E-mail: masicaej@wfu.edu or lalande.danielr@gmail.com

http://www.psypress.com/qjep
http://dx.doi.org/10.1080/17470218.2012.711335

favour of the hypothesis that a relationship or effect exists.

Although the primary emphasis in psychology is to publish results on the basis of NHST (Cumming et al., 2007; Rosenthal, 1979), the use of NHST has long been controversial. Numerous researchers have argued that reliance on NHST is counterproductive, due in large part because *p* values fail to convey such useful information as effect size and likelihood of replication (Clark, 1963; Cumming, 2008; Killeen, 2005; Kline, 2009; Rozeboom, 1960). Indeed, some have argued that NHST has severely impeded scientific progress (Cohen, 1994; Schmidt, 1996) and has confused interpretations of clinical trials (Cicchetti et al., 2011; Ocana & Tannock, 2011). Some researchers have stated that it is important to use multiple, converging tests alongside NHST, including effect sizes and confidence intervals (Hubbard & Lindsay, 2008; Schmidt, 1996). Others still have called for NHST to be completely abandoned (e.g., Carver, 1978).

The goal of the present work was not to resolve the debate surrounding NHST, which has been discussed at length elsewhere, nor was it to argue for a specific view of NHST's appropriateness. Thus, we do not argue that NHST should be abandoned, that it should be supplemented with other tests, or that it should continue to be the main method of statistical inference—each of which are views that have been endorsed elsewhere. Instead, our aim in the current paper was to contribute a new consideration to all sides of the debate. Namely, the aim was to test for evidence of an overreliance on (i.e., a potential misuse of) significance testing, whatever the implications of that may be.

The present paper tested for evidence of an overreliance on significance testing in research publication practices. It did so not by asking researchers to report their confidence in effects linked to various values of *p*, as has been the case in prior work (e.g., Poitevineau & Lecoutre, 2001; Rosenthal & Gaito, 1963). Instead, it examined the psychology literature directly. We examined the distribution of *p* values in a subset of the psychology literature. We expected that if researchers show considerable bias based on NHST standards, then that may be reflected in the distribution of *p* across published effects. More specifically, the number of *p* values immediately below the arbitrary cut-off of .05 may be much higher than would be expected based on the frequency of *p* in other segments of the distribution.

TESTING FOR EVIDENCE

The distribution of *p*

Our approach was to gather all *p* values in a subset of the psychology literature and to assess their distribution. The actual distribution of *p* values in the literature has not been examined in prior research. However, some theoretical papers offer insight into a likely distribution. Sellke, Bayarri, and Berger (2001) simulated *p* value distributions for various hypothetical effects and found that smaller *p* values were more likely than larger ones. Cumming (2008) likewise simulated large numbers of experiments so as to observe the various expected distributions of *p*. To be sure, Cumming's distributions represented *p* values from repeated tests of fixed effects. Still, he found across a wide range of conditions (e.g., a range of effect sizes) that *p* value distributions conformed invariably to exponential curves, with lower values of *p* occurring more frequently than higher ones. Because such curves were observed regardless of the conditions that Cumming imposed, it is reasonable to expect that a sufficiently large sample of *p* values from the literature (drawn from a sufficiently large set of studies testing a range of effect sizes) would also conform to an exponential curve. We examined the literature for such a distribution. Furthermore, we examined whether that distribution was disturbed around the critical value of .05.

Method

Values of *p* were collected from three prominent journals in psychology: *Journal of Experimental Psychology: General* (JEPG), *Journal of Personality and Social Psychology* (JPSP), and *Psychological*

Science (PS). The August 2008 issues and the 11 issues preceding those were examined, so that a total of 12 issues from each journal were assessed. All p values in each journal article were considered. Our focus was on p values within the range of .01 to .10, which included the critical value of .05 as well as a substantial range of values above and below it. Often, exact p values were not reported and were therefore calculated (e.g., using F values and degrees of freedom). All p values within the range of interest were recorded.

As an initial step, both authors examined the 12 PS issues independently of one another so as to establish reliability. Agreement between raters was high so that the sets of p values from 96.8% of articles were identical. Discrepancies in remaining values were attributed to errors in data entry and use of incorrect values for degrees of freedom, and these were corrected through discussion between raters. Because interrater reliability was high for the first 12 issues, and because the few discrepancies that were initially found had no meaningful effect on the resulting p value distributions, the remaining 24 issues from the other two journals were completed by one or the other rater. A total of 3,627 p values between .01 and .10 were collected from the 36 journal issues. Our analyses of JEPG, JPSP, and PS produced 1,092, 1,760, and 705 p values, respectively. We tested the distribution of p for the three journals combined as well as for each individual journal.

A final step in the procedure was aimed at addressing potential experimenter bias. Because the authors were both aware of the purpose of the study, it was possible that errors were made while gathering data that might have facilitated a confirmatory finding. We therefore tested for replication of any results through an additional, independent rater. A research assistant who was blind to the hypothesis was recruited to gather all p values from the PS issues only, and those data were assessed in a separate and final analysis.

Results

One aim of our analysis was to ensure that any anomaly observed in the distribution of p values was not merely an artefact of the manner in which we assessed the distribution. Therefore, we carried out four separate analyses. In each analysis, we divided the range of interest (.01 to .10) into intervals of equal size. The only difference between the four analyses was the size of the intervals into which the range of p values was divided: .01, .005, .0025, or .00125. Varying the interval size enabled us to determine the exact range in which any anomaly in the distribution was present, and it allowed us to ensure that any anomalies were not an artefact of how we divided the distribution.

Fitting the distributions

For each analysis, we counted the number of p values within the various intervals, resulting in distributions of p. Curve estimation procedures were used to determine the best fit for the resulting distributions. When assessing the distribution of p across the three journals together, an exponential model best fitted the data points regardless of the size of the intervals, all R^2s $> .90$ (see Figure 1). The overall distribution of p values thus conformed to a predictable pattern, which highly resembled the curves found in previous simulations (Cumming, 2008; Sellke et al., 2001).

Exponential curves also tended to predict the distribution of p quite well when analysing the distributions from each individual journal. The sole exception occurred when dividing the p value distribution for PS into intervals of .00125, due perhaps to the low numbers of p values within the relatively small intervals. (Some intervals in that distribution contained zero p values, which was never the case when using larger intervals or when assessing the other two journals.) Fit was lowest for the .00125 range ($R^2_{\text{JEPG}} = .81$, $R^2_{\text{JPSP}} = .85$, $R^2_{\text{PS}} = .36$), followed by the .0025 range ($R^2_{\text{JEPG}} = .90$, $R^2_{\text{JPSP}} = .89$, $R^2_{\text{PS}} = .89$), the .005 range ($R^2_{\text{JEPG}} = .95$, $R^2_{\text{JPSP}} = .95$, $R^2_{\text{PS}} = .96$), and the .01 range ($R^2_{\text{JEPG}} = .97$, $R^2_{\text{JPSP}} = .96$, $R^2_{\text{PS}} = .97$).

Assessing the number of p values just below .05

Additional analyses examined whether the number of p values in some intervals departed from the exponential models more than the number of p

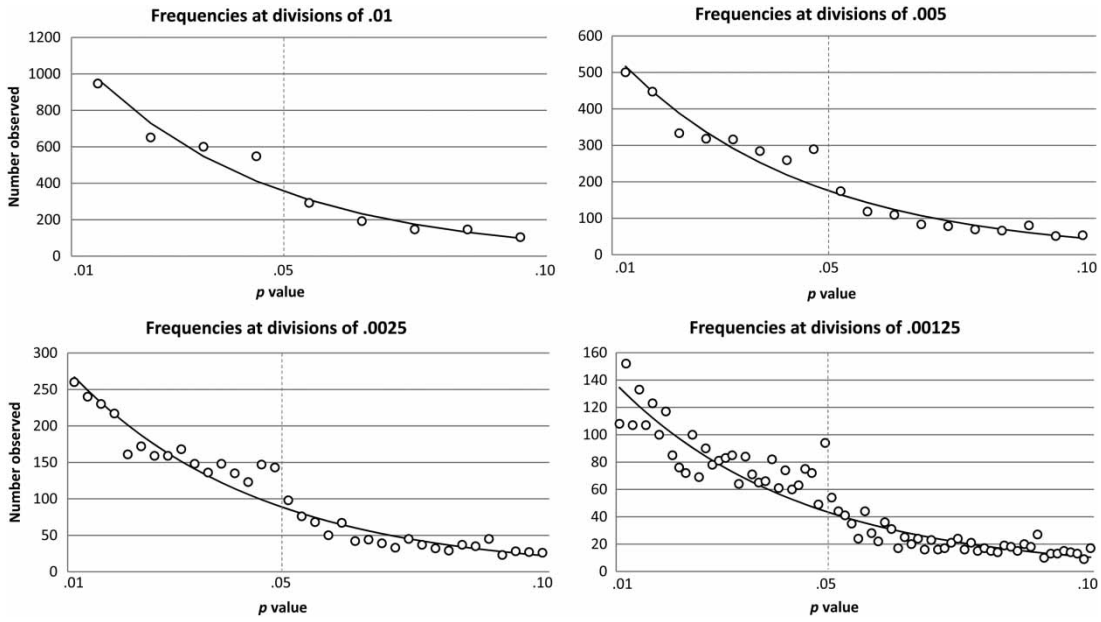

Figure 1.. *The graphs show the distribution of 3,627* p *values from three major psychology journals.*

values in other intervals. For each interval, residuals were obtained by calculating the absolute difference between the observed number of *p* values and the predicted number of *p* values according to the exponential models.

We first assessed the residuals for the distribution of *p* from the three journals combined. Chi-square analyses indicated significant variation in the residuals for all four distributions, $\chi^2(8) = 73.48$, $\chi^2(17) = 98.15$, $\chi^2(35) = 115.68$, $\chi^2(71) = 149.67$, for the .01, 005, .0025, and .00125 models, respectively, all $ps < .0001$. Thus, some residuals within each distribution were significantly larger than others. Moreover, descriptive statistics indicated that the largest residuals in each distribution were found between .045 and .050, and chi-square contrasts (Cox & Key, 1993) confirmed that the residuals for those intervals tended to be significantly larger than were the residuals in other ranges. Descriptive statistics for the residuals and the results of the chi-square contrasts are summarized in Table 1. The results demonstrate that the number of published *p* values that occurred immediately below .05 was much greater than would be expected based on the number of values in the other ranges.

We also assessed the distribution of *p* values for each individual journal. Table 1 displays the descriptive statistics for the residuals and the results of chi-square contrasts that tested whether each residual differed significantly in size from the other residuals in the same distribution. For JEPG, a chi-square analysis revealed only marginally significant variation in the residuals for the .01 model, $\chi^2(8) = 13.9$, $p = .08$. The other models (.005, 0025, and .00125) did not show significant variation in the residuals, $ps > .20$. Nevertheless, the residuals for the intervals between .045 and .050 were either the highest (in the .01 and .005 models) or the second highest in the distributions (in the .0025 and .00125 models; higher residuals were at the extreme end of the distribution where the number of *p* values in each interval were typically small).

For JPSP, all four models revealed significant variation in the residuals, $\chi^2(8) = 44.8$, $\chi^2(17) = 55.9$, $\chi^2(35) = 87.2$, $\chi^2(71) = 165.7$, for the .01, 005, .0025, and .00125 models, respectively, all

Table 1. *Residuals for each range of the various p value distributions*

	Divisions of .01		*Divisions of .005*		*Divisions of .0025*			*Divisions of .00125*				
Journal(s)	.040–.050	All others	.045–.050	All others	.045–.0475	.0475–.050	All others	.045–.04625	.04625–.0475	.0475–.04875	.04875–.05	All others
All three	136.5***	32.4 (23.6)	99.1***	18.7 (14.1)	48.2***	50.9***	10.1 (8.53)	25.9**	24.6**	3.31	49.9***	7.65 (6.51)
JEPG	23.3†	10.7 (6.79)	18.8***	5.86 (4.74)	8.39	11.3†	3.99 (3.11)	1.08	7.58†	5.06	6.52	2.94 (2.47)
JPSP	77.4***	16.1 (15.4)	53.8***	9.46 (8.84)	22.2**	32.4***	6.87 (4.98)	8.97	13.8*	2.41	35.4***	4.51 (3.86)
PS	31.8***	9.30 (12.0)	28.9***	5.05 (7.19)	20.1***	9.55*	3.90 (4.25)	20.2***	2.53	2.88	9.2**	2.83 (3.27)

Note: The table displays the residuals (i.e., the difference between what exponential models predicted and what was actually observed for the number of *p* values) for each interval across the different interval sizes used. Data under "All others" headings represent means with standard deviations in parentheses. Chi-square contrasts (Cox & Key, 1993) were used to test whether each residual differed significantly from the other residuals in the same distribution. JEPG = *Journal of Experimental Psychology: General.* JPSP = *Journal of Personality and Social Psychology.* PS = *Psychological Science.*

†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

ps $< .0001$. In all four models, the highest residuals were for the intervals occurring between .045 and .050.

For PS, all four models also revealed significant variation in the residuals, $\chi^2(8) = 18.7$, $\chi^2(17) = 30.9$, $\chi^2(35) = 57.9$, $\chi^2(71) = 157.6$, for the .01, 005, .0025, and .00125 models, respectively, all ps $< .05$. In all four models, the highest residuals were again for the intervals occurring between .045 and .050.

Results from the blind rater

Final analyses tested whether the occurrence of a large number if p values just below .05 would also be observed in the data collected by a rater who was blind to the study's purpose. The blind rater collected p values from the same set of PS issues as that used previously. Because the disturbance in the p value distribution was found in the previous analyses to be between the values of .045 and .050, the range of p values collected by the blind rater was divided into intervals of .005 only (i.e., the approximate range of the disturbance) rather than the four separate intervals used previously (.01, .005, .0025, and .00125). Once again, the pattern of p value frequencies across the various intervals conformed to an exponential curve ($R^2 = .80$; see Figure 2). Furthermore, the frequencies of p values in some ranges deviated more so from the exponential model than the frequencies in others. A chi-square analysis of the residuals revealed significant variation, $\chi^2(17) = 91.2$, $p < .0001$. In accordance with the previous results, the residual for the .045–.050 range (45.8) was much larger than that of any of the other ranges ($M = 6.57$, $SD = 4.02$).

Discussion

The number of p values in the psychology literature that barely meet the criterion for statistical significance (i.e., that fall just below .05) is unusually large, given the number of p values occurring in other ranges. Specifically, the number of p values between .045 and .050 was higher than that predicted based on the overall distribution of p. This was true when examining all three journals together, and it was true for the distributions of each individual journal. This anomaly is consistent with the proposal that researchers, reviewers, and editors may place undue emphasis on statistical significance to determine the value of scientific results. Biases linked to achieving statistical significance appear to have a measurable impact on the research publication process.

The observed anomaly in the p value distribution could be problematic for a number of reasons. First, the observed skew may be evidence of publication bias. Authors may assume that they must obtain statistical significance in their studies in order to publish (Sterling, 1959). Reviewers and editors likewise may feel pressured to enforce that standard. Second, it exposes an overemphasis on statistical significance, which statisticians have long argued is hurtful to the field (Cohen, 1994; Schmidt, 1996) due in part because p values fail to convey such useful information as effect size and likelihood of replication (Clark, 1963; Cumming, 2008; Killeen, 2005; Kline, 2009; Rozeboom, 1960).

Third, it could be the case that the anomaly is partly a product of researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011). Researchers make many decisions while collecting and analysing their data, and such choices may enable them to nudge research results in a favourable direction. For example, researchers may engage in the "repeated peeks bias" (Sackett, 1979) or "optional stopping" (Wagenmakers, 2007), which amounts essentially to capitalizing on random fluctuations in one's results—one monitors one's data continuously, ceasing only when the desired result appears to have been attained. That and other self-serving practices (e.g., selective exclusion of outliers, selective use of covariates) could be used to disturb p values in a favourable manner. This issue of hidden self-serving biases was touched upon in recent work by Wicherts, Bakker, and Molenaar (2011). They found that researchers were especially unlikely to share their published data for reanalysis if their p values were just below .05. Such a pattern may be indicative of researchers who are reluctant to share their data fearing that erroneous analyses of weak data be exposed.

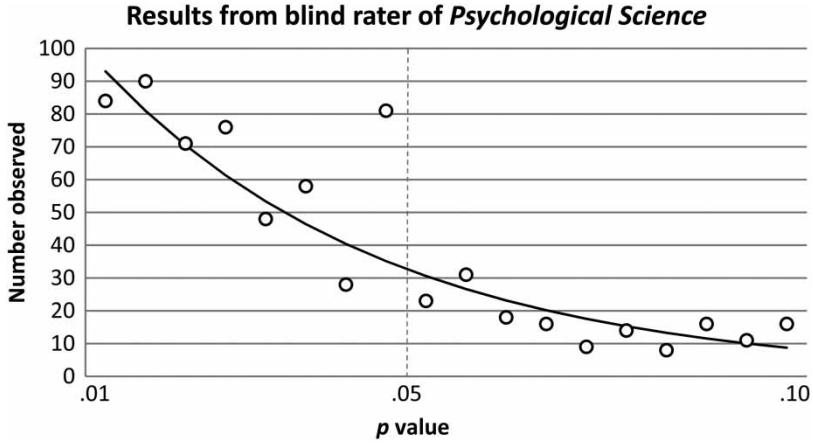

Figure 2. *The distribution of* p *values obtained by the rater who was blind to the hypothesis.*

One limitation of the current work is that we treated the various *p* values as equally independent even though many were from the same journals, authors, papers, and studies. Thus, there were dependencies of various types that we did not control for. Future work may examine to what extent such dependencies may affect the distribution of *p* that was observed. However, while the magnitude of any differences might differ in a multilevel analysis, it seems doubtful that the overall pattern of *p* values and the high frequency of *p*s just below .05 would change.

Future work may probe the precise cause of the observed bump in the *p* value distribution, in order to better address it. If false beliefs about *p* are partly to blame, then one strategy may be to better educate researchers about the proper interpretation of NHST and the benefit of complementary approaches such as likelihood analyses and Bayesian statistics (e.g., Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). If traditions, routines, and norms are partly to blame, then explicit endorsements of alternative data analysis techniques may be needed to counteract the single-minded drive to attain statistical significance. A number of alternative data analysis methods have been proposed (see Killeen, 2005;[1] Kline, 2004, 2009). Finally, if biases linked to the analysis and reporting of data are to blame, then standards for avoiding such self-serving practices may need to be adopted (e.g., Simmons et al., 2011).[1]

To be sure, the implications of the present data will depend on one's view of NHST, of which there are many. Those who endorse the use of NHST may see the present results simply as a warning of its potential misuse. Others have criticized NHST—not because it is or can be used in a biased manner, but because they believe it is a wholly invalid approach (e.g., Cohen, 1994; Schmidt, 1996). For those who endorse this latter view, the present results may serve as further support for arguments either against the use of NHST or at least for the use of complementary analyses.

CONCLUSION

The peer-review publication process determines which scientific findings will inspire future work as well as be dispersed to audiences outside the academic community. The publication process, however, may be perturbed by biases linked to arbitrary standards for evaluating and publishing scientific findings. The present study observed evidence of an overreliance on null-hypothesis significance testing (NHST) in psychological research. NHST

[1]However, see also work by Iverson, Lee, and Wagenmakers (2009) and Trafimow, MacDonald, Rice, and Clason (2010).

may encourage researchers to focus chiefly on achieving a sufficiently low value of *p*. Consistent with that view, the *p* value distribution from three well-respected psychology journals was disturbed such that an unusually high number of *p* values occurred immediately below the threshold for statistical significance. Thus, researchers seem to place undue emphasis on NHST, and the field may benefit from practices aimed at counteracting the single-minded drive toward achieving statistical significance.

Original manuscript received 16 January 2012
Accepted revision received 26 June 2012
First published online 3 August 2012

REFERENCES

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378–399.

Cicchetti, D. V., Koenig, K., Klin, A., Volkmar, F. R., Paul, R., & Sparrow, S. (2011). From Bayes through marginal utility to effect sizes: A guide to understanding the clinical and statistical significance of the results of autism research findings. *Journal of Autism and Developmental Disorders*, *41*(2), 168–174.

Clark, C. A. (1963). Hypothesis testing in relation to statistical methodology. *Review of Educational Research*, *33*, 455–473.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist*, *49*, 997–1003.

Cox, M. K., & Key, C. H. (1993). Post hoc pair-wise comparisons for the chi-square test of homogeneity of proportions. *Educational and Psychological Measurement*, *53*, 951–962.

Cumming, G. (2008). Replication and *p* intervals: p *values predict the future only vaguely, but confidence intervals do much better.* Perspectives on Psychological Science, *3*, 286–300.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., et al. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science*, *18*, 230–232.

Hubbard, R., & Lindsay, R. M. (2008). Why *p* values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology, 18*, 69–88.

Iverson, G. J., Lee, M. D., & Wagenmakers, E.-J. (2009). Prep misestimates the probability of replication. *Psychonomic Bulletin & Review*, *16*(2), 424–429.

Killeen, P. R. (2005). An alternative to null hypothesis significance tests. *Psychological Science*, *16*, 345–353.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in the behavioral sciences*. Washington, DC: American Psychological Association.

Kline, R. B. (2009). *Becoming a behavioral science researcher: A guide to producing research that matters*. New York, NY: The Guilford Press.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301.

Ocana, A., & Tannock, I. F. (2011). When are "positive" clinical trials in oncology truly positive? *Journal of the National Cancer Institute*, *103*, 16–20.

Poitevineau, J., & Lecoutre, B. (2001). Interpretation of significance levels by psychological researchers: The .05 cliff effect may be overstated. *Psychonomic Bulletin & Review*, *8*, 847–850.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, *55*, 33–38.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, *32*, 51–63.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115–129.

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*, 30–34.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights

from Bayes's theorem. *Psychological Review*, *110*, 526–535.

Trafimow, D., MacDonald, J. A., Rice, S., & Clason, D. L. (2010). How often is p_{rep} close to the true replication probability? *Psychological Methods*, *15*(3), 300–307.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE*, *6*(11), e26828.