Research Article

Psychological Science
1–7

DOI: 10.1177/0956797616645672
pss.sagepub.com

Marginally Significant Effects as Evidence for Hypotheses: Changing Attitudes Over Four Decades

Laura Pritschet[1], Derek Powell[2], and Zachary Horne[1]

[1]Department of Psychology, University of Illinois at Urbana–Champaign, and
[2]Department of Psychology, University of California, Los Angeles

Abstract

Some effects are statistically significant. Other effects do not reach the threshold of statistical significance and are sometimes described as "marginally significant" or as "approaching significance." Although the concept of marginal significance is widely deployed in academic psychology, there has been very little systematic examination of psychologists' attitudes toward these effects. Here, we report an observational study in which we investigated psychologists' attitudes concerning marginal significance by examining their language in over 1,500 articles published in top-tier cognitive, developmental, and social psychology journals. We observed a large change over the course of four decades in psychologists' tendency to describe a *p* value as marginally significant, and overall rates of use appear to differ across subfields. We discuss possible explanations for these findings, as well as their implications for psychological research.

Keywords

marginal significance, null-hypothesis significance testing, methodology, open data

Received 1/13/16; Revision accepted 3/31/16

Reflecting on psychologists' use of *p* values, Rosnow and Rosenthal (1989) once remarked, "Surely, God loves the .06 nearly as much as the .05" (p. 1277). God's love notwithstanding, for most psychologists, the .05 alpha criterion conventionally used in null-hypothesis significance testing is the law of the land. Still, many researchers wonder what to do about *p* values that fall near, but do not cross, the threshold of statistical significance.

In the minds of many psychologists, near-threshold *p* values often represent a sort of statistical limbo. Although *p* values greater than .05 have a clear interpretation under the logic of null-hypothesis significance testing—they are nonsignificant *p* values indicating failures to reject the null hypothesis—the .05 criterion is essentially arbitrary. In response to this tension, many researchers label near-threshold *p* values as "marginally significant" or as "approaching significance." Those who take this route carve out a gray area between rejecting and failing to reject the null hypothesis, the precise meaning of which depends on their discretion. A marginal result might be interpreted as a caution against "accepting" the null hypothesis, a promising preliminary result, or sufficient evidence for some noncentral hypothesis, or it might even be interpreted as equivalent to a significant result.

While there are always unwritten elements of the scientific practices of a discipline (e.g., Ariew, 1984; Kuhn, 1962/2012), the statistical criteria used to evaluate findings should not be among them. If the concept of marginal significance is to be used in psychology, then the rules of its use and interpretation should be openly and transparently described. Yet we suspect that any attempt to articulate prescriptions for the use of marginal significance will reveal that this practice is rooted in serious statistical misconceptions. To some extent, these misconceptions are already evident in the mixed statistical

Corresponding Author:
Zachary Horne, University of Illinois at Urbana–Champaign, Department of Psychology, 603 E. Daniel St., Champaign, IL 61820
E-mail: zach.s.horne@gmail.com

heritage of psychology (see Gigerenzer, 2004), which blends the use of the .05 alpha level of Neyman-Pearson decision theory (Neyman & Pearson, 1933)—originally intended to guide decisions between pairs of specific hypotheses—with the null-hypothesis-testing approach of Fisher (1955), who interpreted *p* values as graded evidence against the null. The concept of marginal significance is dubious under either framework. First, the Neyman-Pearson framework is predicated on the use of hard cutoffs that are not meant to be selectively applied. Second, Fisher treated *p* values of .049 and .051 as practically equivalent for the purposes of inference—for Fisher, these *p* values would not lead to meaningfully different conclusions about the likelihood of the data under the null (see Gigerenzer, 2004). Still, most researchers remain concerned with the .05 threshold (e.g., Simonsohn, Nelson, & Simmons, 2014), usually operating as if marginal effects provide less compelling evidence than statistically significant results.

The use of marginal significance is not just a violation of statistical orthodoxy. Researchers often claim that near-threshold *p* values are approaching significance, apparently assuming that the *p* value associated with their statistical test will trend toward zero as data are collected. However, this will be true only if the population effect is nonzero. Thus, this reasoning is circular: Inferring that an effect exists on the basis of a *p* value approaching significance *presumes* that the effect exists, which is, of course, the very question at issue. Yet even experienced psychologists are liable to make this mistake: In a recent study, Yu, Sprenger, Thomas, and Dougherty (2014) found that researchers who monitor their data as it is collected are more likely to collect additional data when a preliminary analysis returns a near-threshold *p* value (e.g., $p = .06$) than when it returns a large *p* value (e.g., $p = .20$). It is now well known that optional stopping of this sort can increase Type I error rates (Simmons, Nelson, & Simonsohn, 2011), but it can also increase Type II error rates. If a preliminary analysis returns a large *p* value, researchers sometimes abandon data collection before achieving an adequately powered sample, thereby increasing Type II errors.

To our knowledge, there are no guidelines—either from top journals or in the American Psychological Association (APA) style guide—prescribing that near-threshold *p* values be labeled or interpreted as marginally significant. This has not always been the case. The second edition of the APA manual (1974) explicitly proscribed the use of marginal significance (although this point has been omitted since the 1983 edition):

> Caution: Do not infer trends from data that fail by a small margin to reach the usual levels of significance. Such results are better interpreted as being caused by chance and are best reported as such. Treat your results section like an income tax return. Take what's coming to you but no more. (p. 19)

Yet, as any experienced reader has surely noticed, the concept of marginal significance has made its way into virtually every empirical journal in the field. This apparent relaxation of the .05 criterion suggests a potentially troublesome state of affairs, as even significant *p* values may constitute weak or inconclusive evidence against the null hypothesis (e.g., Etz & Vandekerckhove, 2016; Good, 1992). The degree to which this practice affects the field is unknown, though there is at least some cause for concern as other seemingly innocuous scientific practices have been shown to reduce the reliability of findings (Simmons et al., 2011). Here, we examined psychologists' tendency to describe results as marginally significant over the last four decades, shedding light on the prevalence of this unwritten statistical practice in the field.

Method

We examined researchers' willingness to describe results as marginally significant in every article published in *Cognitive Psychology* ($n = 98$), *Developmental Psychology* ($n = 564$), and the *Journal of Personality and Social Psychology* (*JPSP*; $n = 873$) in the years 1970, 1980, 1990, 2000, and 2010, which yielded data from 1,535 published articles. These journals were chosen to represent three major subfields of psychology: cognitive, developmental, and social. Further, each journal is among the most prestigious in its subfield, and each has been in continual publication throughout these four decades. Although we did not conduct a power analysis, a decision was made prior to data collection to examine the first year of each decade between 1970 and 2010, with the end goal of providing a snapshot of psychological research practices in contemporary academic psychology. The data reported here can be downloaded from the Open Science Framework (https://osf.io/92xqk/).

Published articles were queried by L. Pritschet, who downloaded the articles using ProQuest from the University of Illinois library system. Articles were searched using Adobe Reader for all instances of the strings "margin" and "approach." L. Pritschet then judged whether these instances were being used to label a result as marginally significant. Table 1 provides examples of phrases coded as indicating marginal significance. If neither "margin" nor "approach" appeared in an article, searches for common words (e.g., "the" and "an") were used to confirm the file was searchable. All documents were searchable and were analyzed in this way. Following the initial coding, a second coder recoded 25% of the data, agreeing on 96.2% of the original codings. Coders then discussed

Table 1. Examples of Sentences From Our Data Set That Indicated Marginal Significance

"Subjects in the sounds condition reported a marginally significant drop in symptom-reporting relative to the control subjects, $t(53) = 1.66$, $p = .10$."
"Girls were found to be more sociable than boys to the mother in the probe and to a marginal extent ($p < .10$) to the stranger in the observations, but only at some of the ages assessed."
"Although the pattern of the means supported the hypothesis, the predicted interaction was only marginally significant, $F(1, 276) = 3.56$, $p < .06$."
"The model predictions were significantly related to observed minority influence, although the [Social Influence Model] predictions were only marginally significant."
"Appraisals of probability differed significantly by emotion in Experience 2 and approached significance in Experience 1, $F(15, 145) = 1.53$, $p = .10$; and appraisals of legitimacy differed significantly by emotion in Experience 1 and approached significance in Experience 2, $F(15, 143) = 1.63$, $p = .07$."
"In summary, of the 18 terms tested to examine the direct or moderating impact of sex on [Iowa Gambling Task] performance, only one (males playing more often on advantageous decks) explained significant variance in the outcome, and only three more approached significance."
"In contrast to looking times, the cardiac data were only marginally supportive of this hypothesis."
"[Adult Attachment Interview] security was positively associated with observed relationship functioning at [Time 1] (deactivation was marginally positively associated with observed functioning; see Table 2)."
"Although the main effect of exposure in the one-way [analysis of variance] only approached significance, $F(3, 39) = 2.27$, $p < .10$, changes in preference between specific preference assessments were significant."

Note: These examples were drawn at random from the data set. The complete list of articles from which these samples are taken is available on the Open Science Framework (https://osf.io/92xqk/).

their disagreements and decided on a final code for each article.

Results

Of the 1,535 articles examined, 66 were theoretical, methodological, review, or commentary articles that did not report experiments, studies, or inferential statistics testing hypotheses (e.g., one such article was a tutorial on using Bayes factors for hypothesis testing). These articles were excluded from subsequent analyses.

First, we examined the unwritten rules researchers follow when labeling p values as marginally significant. Sampling the first p value labeled as marginally significant from each article ($n = 459$) revealed that the vast majority (92.6%) of marginal p values fall between .05 and .10. However, perhaps as a result of the unwritten nature of this rule, p values as large as .18 are sometimes described as marginally significant. The distribution of p values is shown in Figure 1.

Next, we examined researchers' willingness to describe a result as marginally significant over the course of the last four decades and in the three subfields. Figure 2 shows the percentage of articles in which at least one p value was labeled as marginally significant from 1970 to 2010 in each of the three subfields. There are marked differences both across time and among subfields, with an increasing percentage of articles labeling results as marginally significant in later years and more results labeled as marginally significant in social psychology than in developmental and cognitive psychology.

Articles published in 2010 were 2.47 times more likely (95% confidence interval, or CI = [1.88, 3.22]; odds ratio = 3.61) to describe a result as marginally significant than articles published in 1970. Averaging across years, articles published in *JPSP* were 1.60 times more likely (95% CI = [1.35, 1.88];

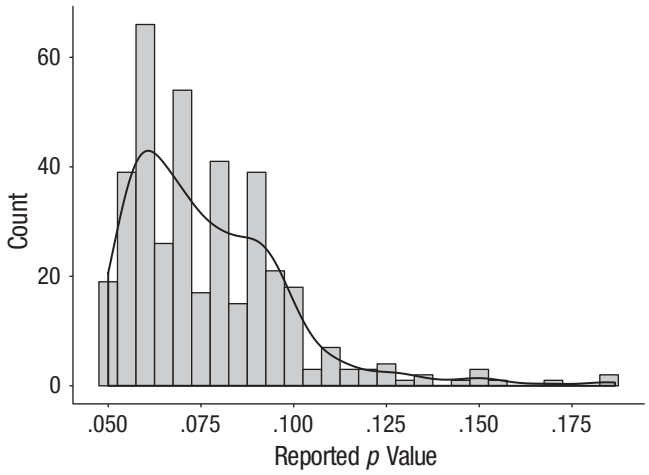

Fig. 1. Histogram of p values labeled as marginally significant in the articles analyzed ($n = 459$). When authors did not report exact p values, we calculated them when sufficient information was available. If sufficient information was not available, p values were estimated using whatever threshold was reported (e.g., $p < .07$ was estimated as .07), except in the case of the "conventional" $p < .10$ threshold. Because we could not accurately estimate the latter, they are excluded from the histogram. Spikes at .06, .07, .08, and .09 are due to original authors' rounding or reporting using thresholds. A kernel density plot is overlaid on the histogram to more clearly show the distribution. One p value below .05 is not shown.

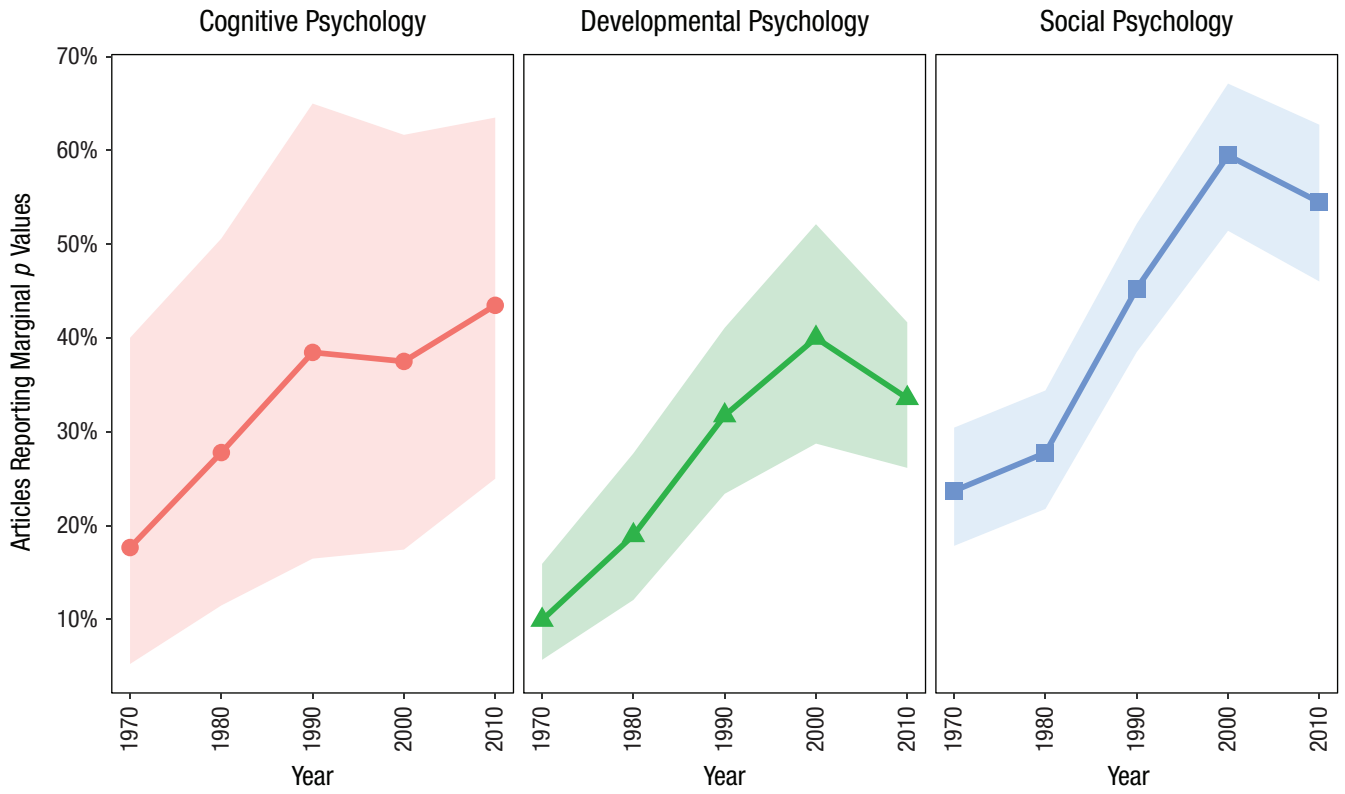

Fig. 2. Percentage of articles in which at least one *p* value was described as marginally significant in each of the 5 years and three subfields analyzed. Shaded areas represent 95% Bayesian credible intervals (Brown, Cai, & DasGupta, 2001).

odds ratio = 2.01) to describe a result as marginally significant than articles published in *Developmental Psychology*. Similarly, articles published in *JPSP* were 1.22 times more likely (95% CI = [0.90, 1.67]; odds ratio = 1.38) to describe a result as marginally significant than articles published in *Cognitive Psychology*. Finally, articles published in *Cognitive Psychology* were 1.30 times more likely (95% CI = [0.94, 1.81]; odds ratio = 1.45) to label results as marginally significant than articles published in *Developmental Psychology*.

We performed logistic regression analyses to evaluate potential differences among subfields and changes across years (summarized in Table 2). These analyses revealed significant differences among subfields—Model A1: $\chi^2(2) = 34.37$, $p < .001$. Adding year of publication to the model revealed a significant effect over and above subfield differences—Model A2: $\chi^2(1) = 88.94$, $p < .001$. Finally, adding interaction terms failed to reduce deviance—Model A3: $\chi^2(2) = 0.987$, $p = .611$—which suggests that these changes across time have occurred in similar fashion among all three subfields.

The change in researchers' tendency to describe results as marginally significant is striking. Whereas the practice was once relatively rare, researchers now appear to label *p* values as marginally significant almost as a matter of course. In 1970, 18% of articles described at least one *p* value as marginally significant. In contrast, in 2000, psychologists were just as likely as not to engage in this practice, with 52% of articles describing at least one result as marginally significant. In fact, the majority of social psychology articles described at least one *p* value as marginally significant in 2000 (59%) and 2010 (54%).

As we have noted, *p* values falling close to the .05 threshold may already constitute weak evidence against the null hypothesis (e.g., Etz & Vandekerckhove, 2016; Good, 1992). Consequently, researchers' increased reliance on marginal results may further reduce the reliability of psychological research.

Further Analyses

We found that researchers working in different subfields differed in their willingness to label results as marginally significant. This suggests that researchers working in different subfields may have different attitudes toward near-threshold *p* values. On the other hand, it is possible that articles published in *Developmental Psychology* and *Cognitive Psychology*, for example, differ in their scope. Larger articles may be more likely to describe at least one result as marginally significant even if researchers' practices are similar across subfields.

To test this possibility, we tabulated the number of experiments reported in each article and found that they differed across subfields and time, as revealed by a 3 (subfield) × 5 (year) analysis of variance, all *p*s < .001. Therefore, we

Table 2. Results of Logistic Regression Analyses Testing for Effects of Subfield, Year of Publication, and Their Interaction on the Probability of Describing at Least One Result as Marginally Significant

Predictor	Model A1	Model A2	Model A3
Subfield 1	−0.323	−0.452*	25.829
	(0.238)	(0.247)	(33.57)
Subfield 2	−0.696**	−0.800**	12.573
	(0.121)	(0.126)	(17.51)
Year of publication	—	0.038**	0.041**
		(0.004)	(0.006)
Year of Publication × Subfield 1	—	—	−0.013
			(0.017)
Year of Publication × Subfield 2	—	—	−0.007
			(0.009)
Constant	−0.370	−72.25	−82.362
Deviance	1,865	1,776	1,775
Model χ^2	$\chi^2(2) = 34.37$	$\chi^2(3) = 123.31$	$\chi^2(5) = 124.2$

Note: Regression coefficients are given for predictors, with standard errors in parentheses. Subfield 1 and Subfield 2 are dummy codes (cognitive: Subfield 1 = 1, Subfield 2 = 0; developmental: Subfield 1 = 0, Subfield 2 = 1; social: Subfield 1 = 0, Subfield 2 = 0).
*$p < .05$. **$p < .001$.

tested the extent to which the number of experiments reported in each article could explain aspects of our findings. We performed a series of logistic regression analyses (summarized in Table 3), finding that the number of experiments reported in an article significantly predicted whether researchers described a result as marginally significant—Model B1: $\chi^2(1) = 67.65$, $p < .001$. Adding dummy variables coding for subfields, we found that an article's subfield was a significant predictor over and above the number of experiments reported in that article—Model B2: $\chi^2(2) = 18.26$, $p < .001$. Finally, we again found that year was a significant predictor in the model—Model B3: $\chi^2(1) = 55.5$, $p < .001$.

Of course, because the number of experiments reported in each article provides only a rough measure of the true size and scope of a research project, we are reluctant to draw firm conclusions on the basis of these results alone. That said, these results suggest that psychologists' attitudes and statistical practices may differ across subfields and may have changed over the last four decades.

Finally, using Model B3, we calculated probabilities (see Fig. 3) and adjusted odds ratios (AORs) to compare researchers' tendencies to describe results as marginally significant across subfields and years of publication. Even after controlling for the year of publication and the

Table 3. Results of Logistic Regression Analyses Comparing Effects of Number of Experiments, Subfield, and Year of Publication on the Probability of Describing at Least One Result as Marginally Significant

Predictor	Model B1	Model B2	Model B3
Number of experiments	0.319**	0.296**	0.189**
	(0.040)	(0.042)	(0.045)
Subfield 1	—	−0.621*	−0.613*
		(0.251)	(0.252)
Subfield 2	—	−0.470**	−0.628**
		(0.127)	(0.132)
Year of publication	—	—	0.032**
			(0.004)
Constant	−1.248	−1.007	−64.45
Deviance	1,831	1,813	1,758
Model χ^2	$\chi^2(1) = 67.65$	$\chi^2(3) = 85.91$	$\chi^2(4) = 141.41$

Note: Regression coefficients are given for predictors, with standard errors in parentheses. Subfield 1 and Subfield 2 are dummy codes (cognitive: Subfield 1 = 1, Subfield 2 = 0; developmental: Subfield 1 = 0, Subfield 2 = 1; social: Subfield 1 = 0, Subfield 2 = 0).
*$p < .05$. **$p < .001$.

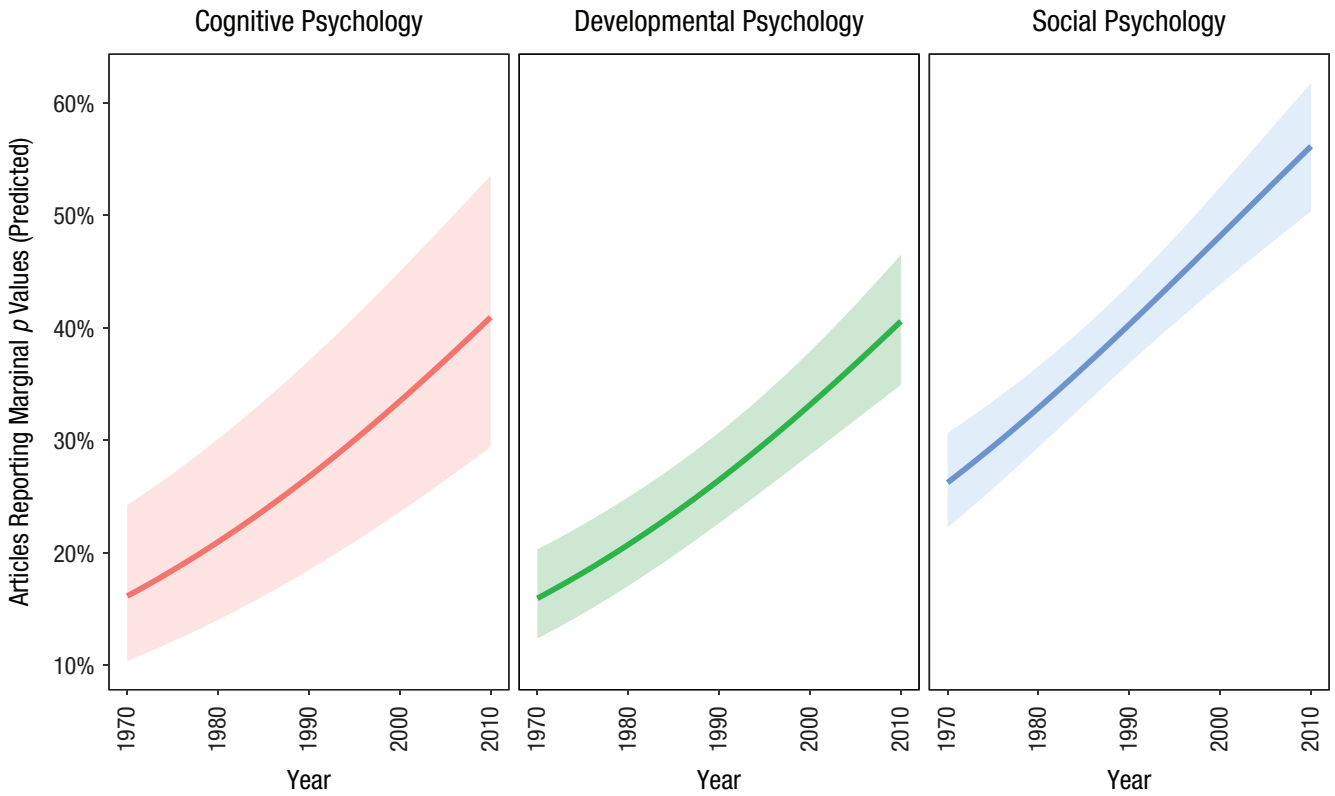

Fig. 3. Estimated percentage of articles in which at least one *p* value would be described as marginally significant in each of the 5 years and three subfields analyzed. Estimates were derived from Model B3 (see Table 2), adjusted for the mean number of experiments per article across subfields (M = 1.89 experiments). Shaded areas represent 95% confidence intervals.

number of experiments reported in each article, we found that articles published in *JPSP* were more likely to describe *p* values as marginally significant than were articles in either *Developmental Psychology* (AOR = 1.87) or *Cognitive Psychology* (AOR = 1.85). In contrast, after adjusting for these other factors, we found that articles published in *Cognitive Psychology* were equally likely to describe a *p* value as marginally significant as were articles in *Developmental Psychology* (AOR = 0.99), which suggests that the differences between developmental and cognitive psychology could be explained by the differences in the number of experiments that researchers in these fields tend to report. Finally, after controlling for the subfield and number of experiments reported, we found that the odds of an article describing a *p* value as marginally significant in 2010 were 3.60 times those of an article published in 1970, consistent with our initial findings.

General Discussion

We observed a large increase in the proportion of articles describing *p* values as marginally significant over the last four decades as well as differences in this practice among three subfields in psychology. It is not immediately clear what our findings say about the field. Is the increased acceptance of marginally significant effects representative of a graded, Fisherian interpretation of *p* values, according to which hard cutoffs are thought to be arbitrary? Or might it suggest the emergence of a more questionable state of affairs for psychological methodology?

In the last decade, there has been an increased emphasis on effect sizes (Cumming, 2013) and Bayesian statistical methods (e.g., Gelman, Carlin, Stern, & Rubin, 2014; Kruschke, 2014). Although some estimation techniques do away with *p* values altogether (e.g., Bayesian estimation), these statistical developments may have encouraged researchers who still rely on *p* values to be less reliant on strict thresholds and to instead view empirical evidence in a more graded fashion. In light of this shifting emphasis in the field as a whole, it is possible that our results reflect that psychologists are more willing to view near-threshold *p* values as evidentially equivalent to their statistically significant counterparts.

Still, less positive methodological changes may be at work. It is well known that there are serious problems with current psychological research practices: Psychologists frequently run underpowered studies (e.g., Cohen, 1992), report engaging in "questionable research practices" (e.g., John, Loewenstein, & Prelec, 2012), and often appear to make ad hoc analysis decisions that can inflate Type I errors (e.g., Gelman & Loken, 2013; Simmons et al., 2011; Yu et al., 2014). Further, methodologists have noted that particular articles, and even entire literatures, appear to be *p-hacked*—flexibly analyzed in order to make the *p* value of

a substantive test less than .05 (e.g., Lakens, 2014). Consistent with these problematic research practices, many findings published in top psychology journals cannot be replicated (Open Science Collaboration, 2015). Indeed, replication concerns may be particularly pressing in social psychology, the subfield in which we observed the largest proportion of articles labeling *p* values as marginally significant.

Taken together, these phenomena may suggest that researchers' increased willingness to describe marginally significant effects as evidence for hypotheses owes to a tacit relaxation of the criterion employed to control the Type I error rate, which may lead to an increased prevalence of findings that provide weak evidence, at best, against the null hypothesis.

Conclusion

Scientists must always be considerate and critical of their research practices. We suspect this point is uncontroversial, but we think that this consideration is particularly important whenever there is reason to believe that standards and practices have changed or may differ among researchers across subfields. It appears that statistical standards have indeed changed, echoing calls for the critical evaluation of the statistical practices used in psychological science.

Action Editor

Hal Arkes served as action editor for this article.

Author Contributions

All of the authors contributed equally to this study. L. Pritschet, D. Powell, and Z. Horne planned the study and collected the data. D. Powell and Z. Horne analyzed the data. L. Pritschet, D. Powell, and Z. Horne wrote the manuscript.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Open Practices

All data have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/92xqk/. The complete Open Practices Disclosure for this article can be found at http://pss.sagepub.com/content/by/supplemental-data. This article has received the badge for Open Data. More information about the Open Practices badges can be found at https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/ and http://pss.sagepub.com/content/25/1/3.full.

References

American Psychological Association. (1974). *Publication Manual of the American Psychological Association* (2nd ed.). Baltimore, MD: Garamond/Pridemark Press.

Ariew, R. (1984). The Duhem thesis. *The British Journal for the Philosophy of Science*, *35*, 313–325.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*, 101–117.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cumming, G. (2013). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, *11*(2), Article e0149794.

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B (Methodological)*, *17*, 69–78.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). London, England: Chapman & Hall/CRC.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Unpublished manuscript. Department of Statistics, Columbia University in the City of New York. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587–606.

Good, I. J. (1992). The Bayes/non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, *87*, 597–606.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. London, England: Academic Press.

Kuhn, T. S. (2012). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press. (Original work published 1962)

Lakens, D. (2014). Grounding social embodiment. *Social Cognition*, *32*(Suppl.), 168–183.

Neyman, J., & Pearson, E. S. (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society*, *29*, 492–510.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 943. doi:10.1126/science.aac4716

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284.

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.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547.

Yu, C. E., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review*, *21*, 268–282.