TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 0 (2021) 1–12
© 2021 Cognitive Science Society LLC
ISSN: 1756-8765 online
DOI: 10.1111/tops.12589

This article is part of the topic "Best of Papers from the 2021 Cognitive Science Society Conference," Andrea Bender (Topic Editor).

Categorical Perception of *p*-Values

V. N. Vimal Rao,[a] Jeffrey K. Bye,[a] Sashank Varma[b]

[a]*Department of Educational Psychology, University of Minnesota*
[b]*School of Interactive Computing, School of Psychology, Georgia Institute of Technology*

Received 25 September 2021; received in revised form 23 October 2021; accepted 26 October 2021

Abstract

Traditional statistics instruction emphasizes a .05 significance level for hypothesis tests. Here, we investigate the consequences of this training for researchers' mental representations of probabilities — whether .05 becomes a boundary, that is, a discontinuity of the mental number line, and alters their reasoning about *p*-values. Graduate students with statistical training ($n = 25$) viewed pairs of *p*-values and judged whether they were "similar" or "different." After controlling for several covariates, participants were more likely and faster to judge *p*-values as "different" when they crossed the .05 boundary (e.g., .046 vs. .052) compared to when they did not (e.g., .026 vs. .032). This result suggests a categorical perception-like effect for the processing of *p*-values. It may be a consequence of traditional statistical instruction creating a psychologically real divide between so-called statistical "significance" and "non-significance." Such a distortion is undesirable given modern approaches to statistical reasoning that de-emphasize dichotomizing the *p*-value continuum.

Keywords: Statistics education; Statistical significance; Probabilistic reasoning; Categorical perception; Rational number processing

1. Categorical perception of *p*-values

The phenomenon of *p*-hacking, wherein researchers make self-serving decisions to achieve "attractive" *p*-values, has spurred debate and reflection among researchers, journal editors, and statisticians (Simmons, Nelson, & Simonsohn, 2011; Simonsohn, Nelson, & Simmons,

Correspondence should be sent to Velliyoor Nott Vimal Rao, Department of Educational Psychology, University of Minnesota, 56 E River Road, Rm 250, Minneapolis, MN 55455, USA. E-mail: rao00013@umn.edu

2014; Tramifow, 2014; Wasserstein, Schirm, & Lazar, 2019). Beyond methodological concerns, we consider here the question of whether instruction and exposure to a literature emphasizing a statistical boundary at .05 results in a mental boundary in researchers' processing of the *p*-value continuum. Cognitive science research has investigated how people use categories to divide cognitive representations of other continuous stimuli in the service of learning and communication (Bruner, Goodnow, & Austin, 1956; Gibson, 1969). These categories have consequences. In particular, the categorical perception effect (CPE; Harnad, 1987) is a distortion in the way people perceive exemplars on the same versus different sides of a category boundary, when controlling for their absolute difference (Fleming, Maloney, & Daw, 2013; Notman, Sowden, & Özgen, 2005).

Given the predominance of the .05 boundary in science and calls for its reform, it is important to understand the underlying cognition in using *p*-values and whether researchers show a similar type of CPE-like effect on the *p*-value continuum between 0 and 1. The existence of such an effect would have implications for statistical hypothesis testing and the recalcitrant phenomenon of *p*-hacking, as such a distortion could affect researchers' statistical interpretations and decisions. It would also potentially spur new research on instruction and practice of statistics. In this study, we adapt a paradigm from categorical perception research to explore, for the first time, whether a CPE-like distortion of *p*-values exists for individuals with statistical training at the graduate level.

2. Background

The formal use of *p*-values in statistical testing traces back to the early 20th century and two competing approaches — significance testing and hypothesis testing. Statistical significance testing compares observed evidence to a candidate hypothesis, through which a researcher estimates a *p*-value by comparing a sample statistic to a hypothesized sampling distribution. A sufficiently small *p*-value, historically one below .05 (Fisher, 1925), indicates that either the hypothesis is true and the observed outcome is coincidentally different from the expectations, or the hypothesis is false.

In the hypothesis testing approach, a candidate hypothesis is compared to a family of possible alternatives to generate a set of decision rules to govern researchers' behavior (Neyman & Pearson, 1933). These rules are determined in a manner that balances the probabilities of two kinds of error when choosing between competing hypotheses — a false rejection of the candidate hypothesis (i.e., Type I error) and a false acceptance of the candidate hypothesis (i.e., Type II error).

Together, these practices combined to form the modern practice of null hypothesis significance testing (NHST), whereby a candidate null hypothesis is rejected if and only if *p* is below a predetermined threshold, canonically .05. Since .05 was originally proposed, this artificial boundary has become a gatekeeper to publication (Stang, Poole, & Kuss, 2010; Tramifow, 2014), despite the admitted arbitrariness of its original proposal (Fisher, 1925) and recommendations that NHST should have at most a limited role in statistical inference (Rao, 1992).

Methodological critiques of NHST have been offered nearly continually since the 1930s, primarily focused on the logic of the approach and its utility in conducting statistical inference

(Cohen, 1994). This has led to a recent movement de-emphasizing hypothesis testing and *p*-values, instead emphasizing estimation via effect sizes and confidence intervals (Cumming, 2014).

Here, we offer a cognitive critique of sorts, considering the representational and reasoning consequences of conceptualizing .05 as boundary delineating two categorical results — statistically "significant" and "nonsignificant." To assess whether *p*-value dichotomization has cognitive consequences, we turn to research on categorization, which is fundamental to human inference and perception (Bruner et al., 1956; Murphy, 2002).

CPEs alter perception such that differences between two values on the same side of a boundary are minimized, while differences between two values on opposite sides of a boundary are exaggerated — even when the physical or numerical difference between the pairs is the same (Goldstone & Hendrickson, 2010; Harnad, 2017). The effect of this phantom discontinuity on perception can lead individuals to make suboptimal decisions (Fleming et al., 2013; Notman et al., 2005). Although not all categories alter perception, CPEs have been shown for a variety of stimuli, including speech sounds and phonemes (e.g., MacKain, Best, & Strange, 1981), colors (e.g., Roberson & Davidoff, 2000), facial expressions (e.g., Etcoff & Magee, 1992), and music, pitch, and rhythm (e.g., Schulze, 1989).

CPEs have been under-researched for numerical stimuli. One notable exception is the demonstration of boundary effects for socially significant categories, such as "thousands" and "millions" (Landy, Charlesworth, & Ottmar, 2017). However, CPEs have not yet been investigated for more frequently experienced numbers, including probabilities limited to the range [0, 1]. The ubiquity of categorical effects for physical stimuli developed through frequent experience suggests that there could be CPE-like effects for numbers in this range. Given their training experience and emphasis on the .05 boundary, it is plausible that researchers may demonstrate a CPE-like effect for *p*-values across the .05 boundary, which may in turn distort their mental representation and processing of *p*-values.

Cognitive models for numerical stimuli suggest that the mental representation of natural numbers is continuous (Ansari, Garcia, Lucas, Hamon, & Dhital, 2005; Moyer & Landauer, 1967). Natural numbers are mapped to points on a mental number line (MNL; Dehaene, Dupoux, & Mehler, 1990). This is also true of rational numbers expressed as decimals (Varma & Karl, 2013), so long as those decimals are neither very small nor very large, that is, neither less than .01 nor greater than .99 (Cohen, Ferrell, & Johnson, 2002).

Moreover, the MNL appears to be a logarithmically compressed distortion of the linear continuum of mathematics. The evidence for this comes in part from experiments where people are asked to identify the greater (or lesser) of a pair of numbers. People make faster judgments when the numbers are small versus large (e.g., 1 vs. 2 is faster than 8 vs. 9; Parkman, 1971); this is the *size effect*. They make faster judgments when the distance between numbers is large versus small (e.g., 2 vs. 8 is faster than 3 vs. 5; Moyer & Landauer, 1967); this is the *distance effect*. Finally, they make faster judgments for pairs of multidigit numbers when the digit in each place of the larger number is greater than its counterpart in the smaller number (e.g., 46 vs. 35 is faster than 45 vs. 36; Nuerk, Weger, & Willmes, 2001; Varma & Karl, 2013); this is the *compatibility effect*. There are also discontinuities of the MNL caused by the place-value symbol system for naming numbers. For example, determining the midpoint

between two numbers is slower and less accurate when the tens digits differ (e.g., bisecting 27–35 is harder than bisecting 21–29; Nuerk, Moeller, Klein, Willmes, & Fischer, 2011); this is the *decade-crossing effect*.

The question we consider here is whether traditional statistics instruction produces a discontinuity in the MNL of p-values at .05, after controlling for the effects of size, distance, compatibility, and decade crossing. If this is *not* the case, then consistent with traditional cognitive models of the MNL, participants should judge $p = .048$ and $p = .051$ to be (1) more similar than $p = .018$ and $p = .021$ because of the size effect and (2) more similar than $p = .018$ and $p = .023$ because of the size and distance effects. However, if a CPE-like effect exists for .05, then individuals may judge $p = .048$ and $p = .051$ to be more different than the other pairs above because only in this case do the two p-values cross the putative .05 boundary. Thus, the goal of this study is to identify whether a CPE-like effect exists in the mental representation of p-values around the .05 boundary.

3. Method

Canonical CPE studies include two tasks to establish a CPE — an identification task to determine the precise location of the boundary (typically implicit) and a discrimination task to confirm within-category indiscriminability. As $p < .05$ is explicitly the conventional boundary for statistical significance, it was assumed to be the location of the boundary separating p-values that would be labeled "statistically significant" or not. To evaluate within-category indiscriminability, we employed the AX discrimination task, which asks participants to identify whether pairs of stimuli are "similar" or "different" (e.g., Repp, 1984). Such judgments are neither inherently correct nor incorrect, and thus a within-subjects design was employed to study patterns in participants' selections across a variety of stimuli.

The experiment was designed to distinguish between the competing predictions of cognitive models of the MNL against those of a hypothetical CPE-like effect. The hypothesized CPE-like effect suggests that individuals will be more likely to label a pair of stimuli as different when they cross the .05 boundary, relative to when the p-values are either both below the boundary (i.e., both "statistically significant") or both above the boundary (i.e., both "not statistically significant"). Furthermore, for a given distance between two p-values, participants will be faster at responding that they are different if they cross the .05 boundary, relative to when they do not cross the .05 boundary.

3.1. Participants

We recruited graduate students associated with the Educational Psychology Department of a large public university in the Midwestern United States. Eligible participants were those who reported having at least 1 full year of experience with hypothesis tests and p-values through coursework, research, or teaching. An initial screening of 40 respondents identified 27 participants who were eligible; of these, 25 completed the experiment in full. Participants were recruited on a voluntary basis and were not financially compensated.

Table 1
Example and number of unique stimuli by within-pair distance and stimulus type

	Example stimuli			Number of unique stimuli		
Distance	Below	.05 Crossing	Above	Below	.05 Crossing	Above
.002	.039 versus .041	.049 versus .051	.079 versus .081	5	2	4
.003	.029 versus .032	.049 versus .052	.069 versus .072	5	4	4
.004	.017 versus .021	.047 versus .051	.057 versus .061	5	3	5
.005	.028 versus .033	.048 versus .053	.078 versus .083	6	4	5
.006	.036 versus .042	.046 versus .052	.086 versus .092	4	3	3
.007	.027 versus .034	.047 versus .054	.067 versus .074	3	3	3
.008	.034 versus .042	.044 versus .052	.054 versus .062	3	4	3
.009	.016 versus .025	.046 versus .055	.086 versus .095	2	4	3

3.2. Materials

Each stimulus was a pair of p-values (see Table 1). The distance between the p-values was varied in thousandths increments to control for an expected distance effect. Between 9 and 15 stimuli pairs were created for each within-pair distance to improve the measurement precision. The critical variable was which boundary was crossed. For each set of 9 to 15 pairs, 2 to 4 crossed the .05 boundary (".05 Crossing"; e.g., $p = .048$ vs. $p = .051$), 2 to 6 crossed a boundary below .05 ("Below"; e.g., $p = .028$ vs. $p = .031$), and 3 to 5 crossed a boundary above .05 ("Above"; e.g., $p = .068$ vs. $p = .071$), as seen in Table 1. It was necessary to include both Below- and Above-boundary stimuli to control for an expected size effect.

All stimuli pairs were of the form "0.0AB," where A and B were nonzero digits to control for a potential effect of a different number of leading zeros on response time (RT) (Schulze, Schmidt-Nielsen, & Achille, 1991). Each pair of p-values had different digits in the hundredths place to control for a potential hundredths-crossing effect analogous to the decade-crossing and tenths-crossing effects (Nuerk et al., 2011; Varma & Karl, 2013).

Distances within each pair ranged from a minimum of .002 to a maximum of .009 to avoid generating a compatibility effect (Nuerk et al., 2001). There were multiple pairs for each distance larger than .002 (e.g., for a distance of .003, the pairs included .049 vs. .052 and .048 vs. .051). We included 18 additional filler stimuli so that participants would not notice patterns in the stimulus set and modify their performance accordingly. The filler stimuli included distances up to .016 and p-values with the same digit in the hundredths place (e.g., $p = .042$ vs. $p = .056$ and $p = .036$ vs. $p = .038$). A total of 108 experimental and filler stimuli were created (see Table 1).

3.3. Procedure

Pairs of p-values were presented sequentially to participants, with the first p-value shown for 1000 ms and the second p-value shown until either a response was made or 5000 ms

elapsed. Participants were asked to "identify whether the *p*-values are similar or different," and to indicate their choice as quickly as possible by pressing either "F" (for similar) or "J" (for different) on their keyboard. In approximately half of the trials, the first *p*-value presented was smaller than the second one, while in the other half it was larger. To induce a statistical mindset, *p*-values were presented in the form "$p = .0AB$." Stimuli were blocked into six sets of 18 pairs, with a 20 seconds break between each block. An additional six stimuli were presented in an initial training phase to familiarize participants with the task procedures.

3.4. Statistical methods

We looked for a CPE-like distortion in two ways. First, we investigated whether "similar" versus "different" judgments (for which there is no objectively correct response) changed as a function of whether the *p*-values crossed the .05 boundary using a log-binomial mixed effects model (e.g., Huang, 2019). Second, we looked for differences in participants' RTs when selecting "different" as a function of the .05 boundary-crossing using a lognormal mixed effects model (e.g., van der Linden, 2006).

To isolate the effect of .05 boundary-crossing, both models adjusted for the size of the *p*-values, the distance between them, and whether the first *p*-value presented was smaller than the second *p*-value presented (or vice versa). Each of these effects were entered into the models as random effects varying across participants to control for possible differences in participants' subjective interpretations of "similar" and "different." Random effects were also included for the order in which stimuli were presented. All filler stimuli were excluded from model fitting.

Preliminary analyses detected aberrant patterns for within-pair distances of .002: participants judged these pairs of *p*-values as different more often than when the distance between the *p*-values was .003 or .004. Because all stimuli pairs with distances of .002 were necessarily of the form .0A9 versus .0B1, we suspect participants might have noticed the particularly salient .049 versus .051 comparison and adjusted their behavior to this .05 crossing stimulus (see Discussion). Additionally, two participants' RTs exhibited exceptionally large RTs, with an average RT over 3000 ms, whereas average RTs for all other participants were below 1500 ms. We suspect that these participants did not attempt to respond as quickly as possible for all stimuli. Due to these suspected response processes compromising data quality, only stimuli with within-pair distances between .003 and .009 were included in the statistical models, and the two participants with large RTs were excluded from the analyses involving RTs.

Taking as our null hypothesis the predictions of cognitive models of the MNL (i.e., of no CPE-like boundary at .05), we used ANOVA Type III sums of squares tests to generate *p*-values for fixed-effects in the log-binomial mixed effects model (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017) and *t*-tests with Satterthwaite's method for degrees of freedom for the lognormal mixed effects model. All analyses were completed using R (R Core Team, 2019) and the lme4 package (Bates et al., 2015).

4. Results

After adjusting for the other potential effects, participants' judgments of similarity versus difference showed the predicted CPE-like effect (see Fig. 1): for a given distance, participants were more likely to label the stimulus pair as "different" for the .05 Crossing stimuli compared to the Above or Below stimuli ($p < .001$; see Table 2). The estimated rate-ratio was 2.42 (95% CI: 1.49–3.95), implying that for a given distance between a pair of p-values, participants were nearly two and a half times as likely to indicate that a stimulus pair was different for the .05 Crossing stimuli, when the p-values were on different sides of the boundary, compared to the Below and Above stimuli, when they were on the same side.

With regard to the covariates, participants were more likely to indicate that a stimulus pair was different as the within-pair distance increased ($p < .001$). There was no effect of size ($p = .240$) nor of presentation order ($p = .484$). Finally, there were no interaction effects

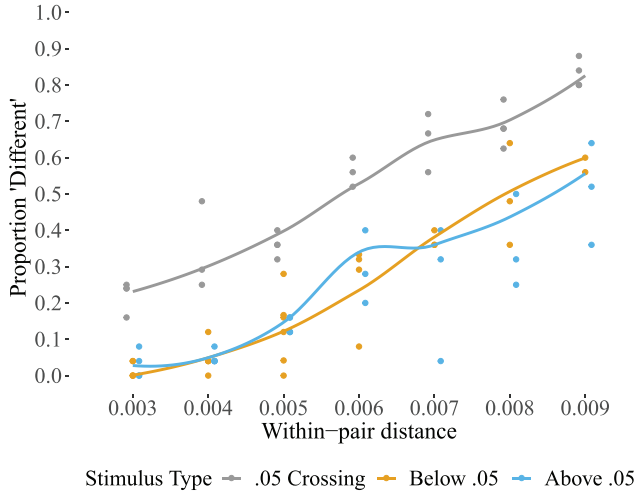

Fig. 1. Unadjusted proportion of participants selecting "different" per stimulus pair (dot) by within-pair distance (x-axis) and stimulus type (color), with weighted LOESS line.

Table 2
Fitted statistical model estimated fixed effects

	Difference selections		Log RT	
Factor	Estimated rate-ratio (95% CI)	p	Estimated RT-ratio (95% CI)	p
.05 Crossing	2.422 (1.49–3.95)	< .001	0.781 (0.62–0.99)	.044
Distance[a]	1.368 (1.28–1.46)	< .001	0.968 (0.94–1.00)	.051
Size[a]	0.991 (0.98–1.01)	.240	1.025 (0.92–1.14)	.750
Smaller first	0.928 (0.75–1.14)	.484	1.001 (0.99–1.00)	.654

[a]Estimates are per 0.001 increase.

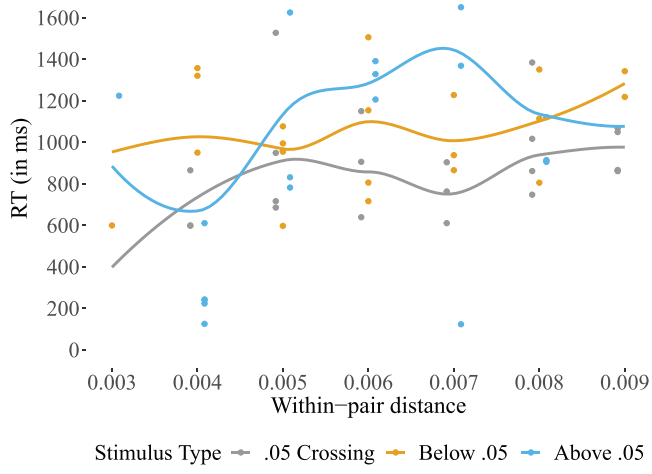

Fig. 2. Unadjusted mean RT for participants' "different" selections per stimulus pair (dot) by within-pair distance (x-axis) and stimulus type (color), with weighted LOESS line.

between any of the factors in a full model, and thus interaction terms were not included in the final model.

The finding of a CPE-like effect at the group level also held at the level of individual participants. Estimates of random effects indicated that 23 of 25 participants were more likely to judge a stimulus pair as "different" when the p-values crossed the .05 boundary, indicating a robust effect. Individuals' estimated rate-ratios ranged from .92 times as likely to select "different" to over 17 times as likely, with a median effect of 1.96 times as likely. These individual differences were not related to whether participants had taken a statistics class within the last year, a proxy for experience and recent exposure to the .05 boundary (Wilcoxon rank-sum test: $p = .768$).

Analysis of participants' RTs similarly showed the predicted CPE-like effect (see Fig. 2): for a given distance, participants were faster to judge a stimulus pair as "different" when the pair crossed the .05 boundary compared to when both p-values were either above or below the boundary ($p = .044$; see Table 2). The estimated percentage reduction in RT was 21.9% (95% CI: 0.7–38.5%) for boundary-crossing pairs, corresponding to a 275 ms reduction in RT relative to the model-adjusted mean RT of 1114 ms.

In addition, participants were faster on average at indicating that a stimulus pair was different as the within-pair distance increased ($p = .051$). There was neither an effect of size ($p = .750$) nor of presentation order ($p = .654$). There were no interaction effects between any of the factors in a full model, and thus interaction terms were not included in the final model.

We also looked for a CPE-like effect on RTs at the individual level. After adjusting for the other effects, 17 of 23 participants were estimated to be faster on average at selecting "different" for the .05 Crossing stimuli. Estimated average percentage differences in RT between

.05 Crossing stimuli and the Above/Below stimuli ranged from 79% faster to 43% slower, with a median of 18% faster. These individual differences were also not related to whether participants had taken a statistics class within the last year (Wilcoxon rank-sum test: $p = .446$).

5. Discussion

This study investigated the mental representation of the *p*-value continuum, specifically whether there is a CPE-like distortion at the .05 boundary. In fact, there was such an effect. Participants were more likely and faster to judge a pair of *p*-values as "different" (vs. "similar") when they crossed the .05 boundary. These effects were present even after controlling for the compatibility, size, distance, and order effects that have been documented in the mathematical cognition literature for comparisons of numbers.

The finding of a CPE-like effect suggests that the mental representation of the *p*-value continuum contains a discontinuity at .05. We hypothesize that this is a consequence of traditional statistics instruction that most graduate students in psychology receive, and also with their reading of a scientific literature still dominated by NHST with a .05 significance level. This hypothesis can be tested in future studies. Specifically, if this instructional and experiential hypothesis is true, then the CPE-like effect for *p*-values should be relatively small for first-year graduate students, should increase over graduate training and statistical coursework, and should be relatively large for working scientists. We will test this prediction this prediction in a follow-up study that includes both a group of graduate students and a group of working scientists.

The current experiment has several limitations. Perhaps most significantly, it included only 25 participants who made only 108 judgments each. Thus, the data were noisy. In addition, the current experiment used only one task, the AX paradigm, to look for a CPE-like effect for *p*-values. This is a simple discrimination task, and participants may have variedly interpreted the choice of "similar" and "different" especially in the absence of a normatively correct response (Gerrits & Schouten, 2004). This subjectivity may have produced the aberrant pattern where participants identified *p*-values with a distance of .002 as different more often than for pairs where the distance was .003 or .004. This aberrant pattern was neither predicted by traditional mathematical cognition theories of the MNL nor a simple CPE distortion of the MNL. We controlled for this subjectivity by including random effects for participants in the statistical models, and more drastically, by limiting the data analysis to stimuli with within-pair distances between .003 and .009. To address these limitations, we will replicate the current study with a larger sample of graduate students and using multiple tasks to look for CPE-like effects for *p*-values.

An open question is whether .05 is a natural boundary in the mental representation of decimals for all individuals regardless of whether they have had statistical training or exposure to the scientific literature. Mathematical cognition research shows that there are "benchmark numbers" for different classes, for example, 0 for integers and $\frac{1}{2}$ for rational numbers (Obersteiner, Alibali, & Marupudi, 2020; Patel & Varma, 2018; Varma & Schwartz, 2011).

It is possible that .05 is a benchmark number for probabilities even outside the *p*-value context. The current study cannot rule out this possible explanation for the current results. Future research should directly contrast this explanation against the instruction-and-experience hypothesis. We will address this question in a follow-up study that includes both a group of statistically untrained undergraduates and a group of statistically trained graduate students.

An interesting question is whether the *p*-value continuum is partitioned into more categories than "statistically significant" and "nonsignificant." Additional labels, such as "marginally significant" and "nearly significant" are common in the psychological literature (Pritschet, Powell, & Horne, 2016). These additional labels may have induced a complex categorization of *p*-values into trichotomies or tetrachotomies. Further studies are needed to investigate whether these categories are also psychologically real to graduate students and working scientists in psychological science.

Recently, statisticians have argued against the categorization of the *p*-value continuum into statistical significance and nonsignificance (Wasserstein et al., 2019). The advice is meant to obviate the phenomenon of *p*-hacking and to quell the NHST controversy. However, categorizations are generally helpful to novice learners (Gibson, 1969), and some statisticians have pushed back on the abandonment of statistical significance in the classroom for this reason (e.g., Krueger & Heck, 2019).

The current study provides an initial investigation into the mental representations underlying the interpretation of *p*-values. The results show that CPE-like effects for *p*-values may exist for emerging psychological scientists. This might be taken as evidence for the dismal conclusion that *p*-hacking and other questionable research practices may be an inevitable consequence of how people think and learn. We do not believe this to be the case. CPEs that result from early categorizations are not permanent — categorizations can change through experience (Goldstone, 1994). From this perspective, our results set the stage for future research on mental representations of *p*-values and their tuning through statistical instruction. An important direction for future research, then, is to develop classroom activities, assessed by instructional studies, that teach NHST to students without distorting their mental representations of probability. This would be an important step toward moving to a world beyond "$p < .05$."

References

Ansari, D., Garcia, N., Lucas, E., Hamon, K., & Dhital, B. (2005). Neural correlates of symbolic number processing in children and adults. *Neuroreport*, *16*(16), 1769–1773. https://doi.org/10.1097/01.wnr.0000183905.23396.f1

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … & Bolker, M. B. (2015). Package 'lme4'. *Convergence*, *12*(1), 2. Retrieved from https://github.com/lme4/lme4/

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley. https://doi.org/10.2307/1292061

Cohen, D. J., Ferrell, J. M., & Johnson, N. (2002). What very small numbers mean. *Journal of Experimental Psychology: General*, *131*(3), 424–442. https://doi.org/10.1037/0096-3445.131.3.424

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 626–641. https://doi.org/10.1037/0096-1523.16.3.626

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, *44*(3), 227–240. https://doi.org/10.1016/0010-0277(92)90002-Y

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver and Boyd.

Fleming, S. M., Maloney, L. T., & Daw, N. D. (2013). The irrationality of categorical perception. *Journal of Neuroscience*, *33*(49), 19060–19070. https://doi.org/10.1523/JNEUROSCI.1263-13.2013

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics*, *66*(3), 363–376. https://doi.org/10.3758/BF03194885

Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York: Appleton-Century-Crofts.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200. https://doi.org/10.1037/0096-3445.123.2.178

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *WIREs: Cognitive Science*, *1*(1), 69–78. https://doi.org/10.1002/wcs.26

Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. In Harnad S., *Categorical perception: The groundwork of cognition* (pp. 1–28). New York: Cambridge University Press.

Harnad, S. (2017). To cognize is to categorize: Cognition is categorization. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (2nd ed., pp. 21–54). Elsevier. https://doi.org/10.1016/B978-0-08-101107-2.00002-6

Huang, F. L. (2019). Alternatives to logistic regression models in experimental studies. *Journal of Experimental Education*, https://doi.org/10.1080/00220973.2019.1699769

Krueger, J. I., & Heck, P. R. (2019). Putting the *p*-value in its place. *American Statistician*, *73*(S1), 122–128. https://doi.org/10.1080/00031305.2018.1470033

Landy, D., Charlesworth, A., & Ottmar, E. (2017). Categories of large numbers in line estimation. *Cognitive Science*, *41*(2), 326–353. https://doi.org/10.1111/cogs.12342

MacKain, K. S., Best, C. T., & Strange, W. (1981). Categorical perception of English /r/and /l/ by Japanese bilinguals. *Applied Psycholinguistics*, *2*(4), 369–390. https://doi.org/10.1017/S0142716400009796

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*(5109), 1519–1520. https://doi.org/10.1038/2151519a0

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/1602.001.0001

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, *231*, 289–337. 694-706https://doi.org/10.1098/rsta.1933.0009

Notman, L. A., Sowden, P. T., & Özgen, E. (2005). The nature of learned categorical perception effects: A psychophysical approach. *Cognition*, *95*(2), B1–B14. https://doi.org/10.1016/j.cognition.2004.07.002

Nuerk, H. C., Moeller, K., Klein, E., Willmes, K., & Fischer, M. H. (2011). Extending the mental number line. *Zeitschrift für Psychologie*, *219*(1), 3–22. https://doi.org/10.1027/2151-2604/a000041

Nuerk, H. C., Weger, U., & Willmes, K. (2001). Decade breaks in the mental number line? Putting the tens and units back in different bins. *Cognition*, *82*(1), B25–B33. https://doi.org/10.1016/S0010-0277(01)00142-1

Obersteiner, A., Alibali, M. W., & Marupudi, V. (2020). Complex fraction comparisons and the natural number bias: The role of benchmarks. *Learning and Instruction*, *67*, 101307. https://doi.org/10.1016/j.learninstruc.2020.101307

Parkman, J. M. (1971). Temporal aspects of digit and letter inequality judgments. *Journal of Experimental Psychology*, *91*(2), 191–205. https://doi.org/10.1037/h0031854

Patel, P. J., & Varma, S. (2018). How the abstract becomes concrete: Irrational numbers are understood relative to natural numbers and perfect squares. *Cognitive Science*, *42*, 1642–1676. https://doi.org/10.1111/cogs.12619

Pritschet, L., Powell, D., & Horne, Z. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science*, *27*(7), 1036–1042. https://doi.org/10.1177/0956797616645672

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, C. R. (1992). R. A. Fisher: The founder of modern statistics. *Statistical Science*, *7*(1), 34–48. https://doi.org/10.1214/ss/1177011442

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. *Speech and Language*, *10*, 243–335. https://doi.org/10.1016/B978-0-12-608610-2.50012-1

Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition*, *28*(6), 977–986. https://doi.org/10.3758/BF03209345

Schulze, H. H. (1989). Categorical perception of rhythmic patterns. *Psychological Research*, *51*(1), 10–15. https://doi.org/10.1007/BF00309270

Schulze, K. G., Schmidt-Nielsen, A., & Achille, L. B. (1991). Comparing three numbers: The effect of number of digits, range, and leading zeros. *Bulletin of the Psychonomic Society*, *29*(4), 361–364. https://doi.org/10.3758/BF03333945

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*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Stang, A., Poole, C., & Kuss, O. (2010). The ongoing tyranny of statistical significance testing in biomedical research. *European Journal of Epidemiology*, *25*(4), 225–230. https://doi.org/10.1007/s10654-010-9440-x

Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology*, *36*(1), 1–2. https://doi.org/10.1080/01973533.2014.865505

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204. https://doi.org/10.3102/10769986031002181

Varma, S., & Karl, S. R. (2013). Understanding decimal proportions: Discrete representations, parallel access, and privileged processing of zero. *Cognitive Psychology*, *66*(3), 283–301. https://doi.org/10.1016/j.cogpsych.2013.01.002

Varma, S., & Schwartz, D. L. (2011). The mental representation of integers: A symbolic to perceptual-magnitude shift. *Cognition*, *121*, 363–385. https://doi.org/10.1016/j.cognition.2011.08.005

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$". *American Statistician*, *73*(S1), 1–19. https://doi.org/10.1080/00031305.2019.1583913