Categorical Perception of *p*-values

V.N. Vimal Rao (rao00013@umn.edu)
Department of Educational Psychology, University of Minnesota
56 E. River Road, Minneapolis, MN 55455 USA

Jeffrey Kramer Bye (jbye@umn.edu)
Department of Educational Psychology, University of Minnesota
56 E. River Road, Minneapolis, MN 55455 USA

Sashank Varma (varma@gatech.edu)
School of Interactive Computing, Georgia Institute of Technology
85 Fifth Street NW, Atlanta, GA 30308 USA

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, i.e., a discontinuity of the mental number line, and alters their perception of differences between *p*-values. Graduate students ($n = 25$) with statistical training viewed pairs of *p*-values and judged whether they were 'similar' or 'different'. After controlling for covariates, participants were more likely and faster to judge *p*-values as 'different' when they crossed the .05 boundary (e.g., .047 vs. .052) compared to when they did not (e.g., .027 vs. .032). This categorical perception effect suggests that traditional statistical instruction creates a psychologically real divide between so-called statistically significant and non-significant *p*-values. Such a distortion is undesirable given modern approaches to statistical reasoning that de-emphasize dichotomizing *p*-values.

Keywords: statistics education; statistical significance; categorical perception; rational number processing.

Introduction

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 et al., 2011; Simonsohn et al., 2014; Tramifow, 2014; Wasserstein et al., 2019). Beyond methodological concerns, we consider here the question of whether instruction and practice emphasizing a statistical boundary at .05 results in a mental boundary in people's understanding and perception of probabilities. Cognitive science research has investigated how people use categories to divide cognitive representations of continuous stimuli in the service of learning and communication (Bruner et al., 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 vs. different sides of a category boundary (Fleming et al., 2013; Notmon et al., 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 CPE on the *p*-value continuum between 0 and 1. The presence of distortions could affect researchers' statistical interpretations and decisions. The potential existence of such distortions would shed new light on statistical hypothesis testing and the recalcitrant phenomenon of *p*-hacking. It would also potentially spur new research on instruction and practice of statistics. In this study, we explore whether a CPE for *p*-values exists in individuals with statistical training at the graduate level.

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 expectations based on a candidate hypothesis to observed evidence. A researcher estimates a *p*-value, or the probability of observing a possible outcome at least as deviant as the observed evidence from the expectation based on the hypothesis. A sufficiently small *p*-value, historically those below .05 (Fisher, 1925), indicates that either the hypothesis is true and the observed outcome is deviant by a rare coincidence 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 minimizes 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 $p < .05$.

Since then, the artificial boundary of .05 has become a gatekeeper to publication (Tramifow, 2014; Stang et al., 2010), despite 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 emphasizing estimation via effect sizes and confidence intervals while de-emphasizing hypothesis testing and *p*-values (Cumming, 2014).

Here, we offer a cognitive critique of sorts, considering the representational and reasoning consequences of conceptualizing .05 as the boundary delineating 'statistically significant' results. We begin with 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; Notmon et al., 2005). Although not all categorizations alter perception, CPEs have been shown for a variety of stimuli including facial expressions (e.g., Etcoff & Magee, 1992), speech sounds and phonemes (e.g., MacKain et al., 1981), colors (e.g., Roberson & Davidoff, 2000), and music, pitch, and rhythm (e.g., Schulze, 1989).

Notably, categorical perception effects have not yet been investigated for numeric stimuli such as probabilities limited to the range [0, 1]. However, boundary effects have been found for socially significant categories such as 'thousands' and 'millions' (Landy et al., 2017). Furthermore, the ubiquity of categorical effects for physical stimuli suggests that there may be a CPE for numbers in the [0, 1] range and that it may in turn distort the mental representation of *p*-values.

Cognitive models for numeric stimuli suggest that the mental representation of natural numbers is continuous (Ansari et al., 2005; Moyer & Landauer, 1967). Natural numbers are mapped to points on a mental number line (MNL; Dehaene et al., 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 (i.e., not less than .01 nor greater than .99; Cohen et al., 2002).

Moreover, the MNL appears to be distorted from the linear continuum of mathematics. The evidence for a logarithmically compressed MNL 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; LeFevre et al., 1996); this is the size effect. They make faster judgments when the distance between numbers is far vs. near (e.g., 2 vs. 8 is faster than 3 vs. 5; Cohen, 2010); this is the distance effect. Finally, they make faster judgments for pairs of multi-digit 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 et al., 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 21 – 29; Nuerk et al., 2011); this is the decade-crossing effect.

The question we consider here is whether traditional statistics instruction produces a discontinuity in the MNL at .05 after controlling for the effects of size, distance, compatibility, and decade crossing. If this is ***not*** the case, consistent with traditional cognitive models of the MNL, participants should perceive $p = .048$ and $p = .051$ to be *more similar* than $p = .018$ and $p = .021$ because of the size effect and *more similar* than $p = .018$ and $p = .023$ because of the size and distance effects. However, if a CPE exists, individuals may perceive $p = .048$ and $p = .051$ to be *more different* than at least some of 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 effect exists in the perception of *p*-values.

Methods

Canonical CPE studies include two tasks to establish a CPE — an identification task to determine the precise location of the boundary and a discrimination task to confirm within-category indiscriminability. As $p < .05$ is the conventional boundary for statistical significance, it was assumed to be the location of the boundary separating *p*-values that would be labelled 'statistically significant'. To evaluate within-category indiscriminability, we employed the AX discrimination task, which asks participants to identify when 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 the MNL theory against those of a hypothetical CPE — a hypothetical CPE 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 would be faster when responding that they are different if the *p*-values cross the .05 boundary, relative to when then *p*-values do not cross.

Participants

We recruited graduate students associated with the Department of Educational Psychology at the University of Minnesota. Eligible participants were those who reported having experience through research or teaching with hypothesis tests and *p*-values. An initial screening of 40 respondents identified 25 who were eligible and completed the experiment in full. Participants were recruited on a voluntary basis and were not compensated for participation.

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. The critical variable was boundary-crossing. Of these 9-15 pairs, 3-4 crossed the .05 boundary, 3-5 were below the boundary, and 3-5 were above the boundary. 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 non-zero digits to control for a potential effect of a different number of leading zeros on response time (Schulze et al., 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., 2001; Varma & Karl, 2013).

Distances within a pair ranged from .002 to a maximum of .009 to avoid pairs generating a compatibility effect, and there were multiple pairs for each distance larger than .002 (e.g., .049 vs. .052 and .048 vs. .051 for a within-pair distance of .003). We included 18 additional filler stimuli so that participants would not notice and begin attending to patterns in the stimulus set. These filler stimuli included distances up to .016 and *p*-values with the same digit in the hundredths place. A total of 108 stimuli were created (see Table 2).

Table 1: Example stimuli by within-pair distance and stimulus type

Distance	Below*	.05 Crossing
.002	.039 vs. .041	.049 vs. .051
.003	.029 vs. .032	.049 vs. .052
.004	.017 vs. .021	.047 vs. .051
.005	.028 vs. .033	.048 vs. .053
.006	.036 vs. .042	.046 vs. .052
.007	.027 vs. .034	.047 vs. .054
.008	.034 vs. .042	.044 vs. .052
.009	.016 vs. .025	.046 vs. .055

*Above pairs mirrored Below pairs. Each within-pair distance had 9-15 unique pairs.

Table 2: Number of unique stimuli by within-pair distance and stimulus type

Distance	Below	.05 Crossing	Above
.002	5	2	4
.003	5	4	4
.004	5	3	5
.005	6	4	5
.006	4	3	3
.007	3	3	3
.008	3	4	3
.009	2	4	3
filler	6	6	6

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 = 0.0AB". Stimuli were blocked into six sets of 18 pairs, with a 20-second break between each block. An additional six stimuli were presented in an initial training phase to familiarize participants with the task procedures.

Statistical Models

We looked for a CPE in two ways. First, we investigated whether similar vs. 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' response times (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 size of the *p*-values, the distance between them, and whether the first *p*-value presented was smaller than the second *p*-value presented. Each of these effects were entered into the models as random effects varying across participants. Random effects were also included for the order in which stimuli were presented. All filler stimuli were excluded from analysis.

Preliminary analyses detected aberrant patterns for within-pair distances of .002, in which participants judged *p*-values as different more often than when distances between *p*-values were .003 or .004. As all stimuli pairs with distances of .002 were necessarily of the form .0A9 vs. .0B1, we suspect participants might have noticed the particularly salient .049 vs. .051 comparison and adjusted their behavior to this .05 Crossing stimulus (see Discussion). Additionally, two participants' RTs exhibited aberrant patterns of exceptionally large RTs. We suspect that these participants did not attempt to respond as quickly as possible. Due to these suspected response process compromising data quality, only stimuli with within-pair distances between .003 and .009 were included in the statistical models, and two participants' RTs were therefore excluded from analyses of RTs.

Taking as our null hypothesis the MNL theories' predictions, we used ANOVA Type III Sums of Squares tests to generate *p*-values for fixed-effects in the log-binomial mixed effects model (Matuschek et al., 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).

Results

After adjusting for the other potential effects, judgments of similarity vs. difference showed the predicted CPE (see Figure 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 ($p < 0.0001$; see Table 3). 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 the pair were different in .05 Crossing stimuli, compared to when the *p*-values were on the same side of the boundary, as in the Below and Above stimuli.

In addition, participants were more likely to indicate that a stimulus pair was different as the within-pair distance increased ($p < .0001$). There were no effects of size ($p = 0.2404$) nor presentation order ($p = 0.4838$). There were no interaction effects between any of the factors in the model.

The finding of a CPE 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 boundary, indicating a robust effect. 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 (Wilcoxon Rank-Sum Test: $p = 0.7675$).

Analysis of participants' RTs similarly showed the predicted CPE (see Figure 2) — for a given distance, participants were faster to judge a stimulus pair as 'different' when the pair was a .05 Crossing stimulus compared to the Above or Below stimuli ($p = 0.0440$; see Table 4). The estimated percentage reduction in RT was 21.9% (95% CI: 0.7% – 38.5%), corresponding to a 275 ms reduction in RT compared to the model-adjusted mean RT of 1114 ms.

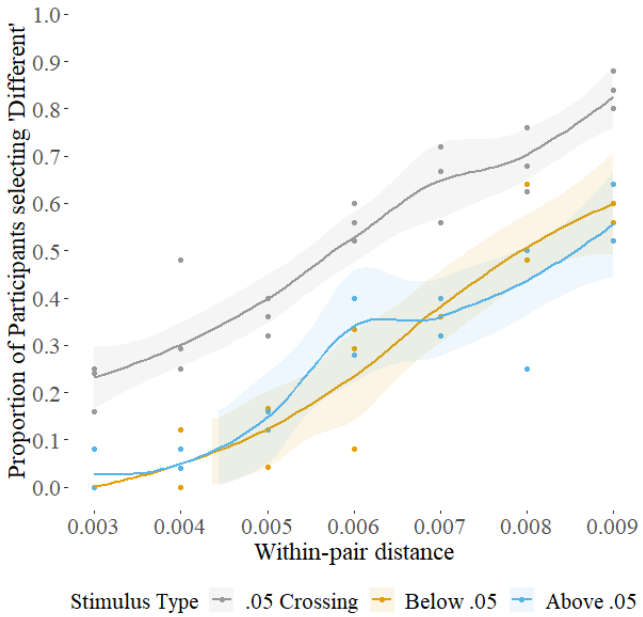

Figure 1: Unadjusted proportion of participants selecting 'Different' by within-pair distance and stimulus type, with 95% Confidence LOESS Envelope.

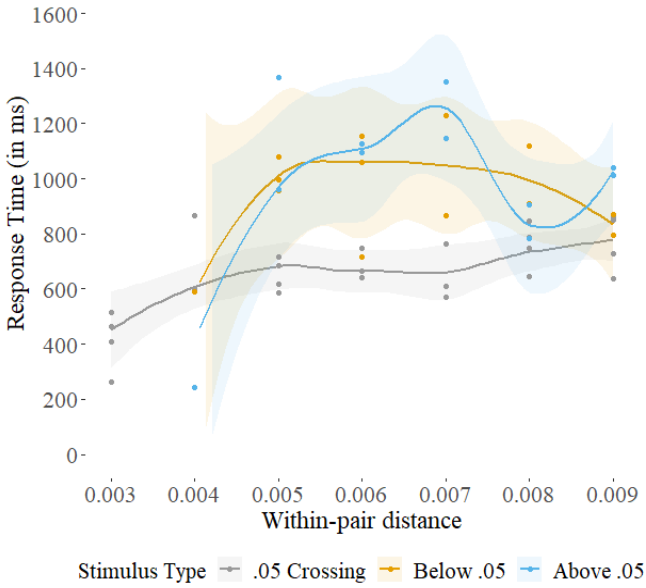

Figure 2: Unadjusted mean RT for participants' difference selections by within-pair distance and stimulus type, with 95% confidence weighted LOESS envelope.

Table 3: Fitted log-binomial mixed effects model predicting participants' difference selections

Factor	Estimated rate-ratio (95% CI)	*p*-value
.05 Crossing	2.422 (1.49 – 3.95)	<0.0001
Distance*	1.368 (1.28 – 1.46)	<0.0001
Size*	0.991 (0.98 – 1.01)	0.2404
Smaller 1st	0.928 (0.75 – 1.14)	0.4838

*Estimates are per 0.001 increase

Table 4: Fitted lognormal mixed effects model predicting participants' log RT when selecting 'different'

Factor	Estimated RT-ratio (95% CI)	*p*-value
.05 Crossing	0.781 (0.62 – 0.99)	0.0440
Distance*	0.968 (0.94 – 1.00)	0.0506
Size*	1.025 (0.92 – 1.14)	0.7504
Smaller 1st	1.001 (0.99 – 1.00)	0.6538

*Estimates are per 0.001 increase

In addition, participants were on average faster at indicating that a stimulus pair was different as the within-pair distance increased ($p = 0.0506$). There were no effects of size ($p = 0.7504$) nor presentation order ($p = 0.6538$). There were no interaction effects between any of the factors in the model.

We also looked for a CPE on RTs at the individual level. After adjusting for the other effects, 17 of 23 participants were on average faster at selecting 'different' in .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 not related to whether participants had taken a statistics class within the last year (Wilcoxon Rank-Sum Test: $p = 0.4458$).

Discussion

This study investigated the mental representation of the *p*-value continuum, specifically whether there is a CPE at the boundary of .05. 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 compatibility, size, distance, and order effects in the comparisons of numbers that have been documented in the mathematical cognition literature.

The finding of a CPE indicates that the *p*-value continuum contains a discontinuity at .05. We hypothesize that this is a consequence of traditional statistics instruction, and with reading a scientific literature still dominated by NHST with $\alpha = .05$. Specifically, if this instructional and experiential hypothesis is true, then the CPE 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.

The current experiment has several limitations. First, it included only 25 participants who made only 108 judgments each. Thus, the data were noisy, and the RT data exhibited unexpected patterns which occluded precise estimation of the CPE and its interaction with distance effects.

Second, the CPE found here might be specific to the AX paradigm used, and more generally the adaptation of CPE tasks to numeric stimuli. The AX task is a simple discrimination task, and participants may have interpreted the choice of 'similar' and 'different' subjectively, especially in the absence of a correct response (Gerrits & Schouten, 2004). This subjectivity may also have interacted with a complex distortion of the MNL to produce the aberrant pattern where participants identified *p*-values with a distance of .002 as different more often than for within-pair distances of .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. In the present study, we have controlled for this subjectivity by including random effects for participants in statistical models and limiting the data analysis to stimuli with within-pair distances between .003 and .009. Additionally, we are currently conducting further studies investigating whether .05 is a psychological boundary across a range of discrimination tasks, examining the structure of the distortion in greater detail, and considering methods to adapt CPE discrimination tasks for numeric stimuli.

Third, although traditional CPE study participants compare two categories with distinctive labels (e.g., 'P' and 'B'), *p*-values' categories are simply 'statistically significant' and '*not* statistically significant'. This asymmetry of categorization suggests there may be complex asymmetries near .05 in the MNL. Further studies are needed to understand the nature of a CPE for complementary categories such as 'statistically significant' and 'not statistically significant'.

Nevertheless, the current study provides a first insight into the mental representations underlying the interpretation of *p*-values and sets the stage for future research on these representations and their tuning through statistical instruction.

Recently, statisticians have suggested that comparisons of statistical significance and non-statistical significance should not be made (Wasserstein et al., 2019). The advice is meant to obviate the phenomenon of *p*-hacking and to quell the NHST controversy. However, dichotomous 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).

Our results show that CPEs 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 and perceptions can change through experience (Goldstone, 1994). Thus, an important direction for future research is to develop classroom activities and run instructional studies showing that NHST can be taught to students without distorting their perceptions of probability. This would be an important step towards moving to a world beyond '$p < 0.05$'.

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.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … & Bolker, M. B. (2015). Package 'lme4'. *Convergence, 12*(1), 2.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Cohen, D. J. (2010). Evidence for direct retrieval of relative quantity information in a quantity judgment task: Decimals, integers, and the role of physical similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1389.

Cohen, D. J., Ferrell, J. M., & Johnson, N. (2002). What very small numbers mean. *Journal of Experimental Psychology: General, 131*(3), 424–442.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*(12), 997–1003.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science, 25*(1), 7-29.

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.

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition, 44*(3), 227-240.

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.

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & psychophysics, 66*(3), 363-376.

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.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *WIREs: Cognitive Science, 1*(1), 69-78.

Harnad, S. (1987) Psychophysical and cognitive aspects of categorical perception: A critical overview. In Harnad, S. (ed.) (1987) *Categorical Perception: The Groundwork of Cognition*. 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.

Huang, F. L. (2019). Alternatives to logistic regression models in experimental studies. *The Journal of Experimental Education*, 1-16.

Krueger, J. I., & Heck, P. R. (2019). Putting the p-value in its place. *The American Statistician, 73*(S1), 122-128.

Landy, D., Charlesworth, A., & Ottmar, E. (2017). Categories of large numbers in line estimation. *Cognitive science, 41*(2), 326-353.

LeFevre, J. A., Sadesky, G. S., & Bisanz, J. (1996). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(1), 216.

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.

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.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature, 215*(5109), 1519-1520.

Murphy, G. L. (2002). *The big book of concepts.* Cambridge, MA: MIT Press.

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. Series A, Containing Papers of a Mathematical or Physical Character, 231*(694-706), 289-337.

Notman, L. A., Sowden, P. T., & Özgen, E. (2005). The nature of learned categorical perception effects: a psychophysical approach. *Cognition, 95*(2), B1-B14.

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.

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.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rao, C. R. (1992). R. A. Fisher: The founder of modern statistics. *Statistical Science, 7*(1), 34-48.

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In *Speech and Language* (Vol. 10, pp. 243-335). Elsevier.

Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition, 28*(6), 977-986.

Schulze, H. H. (1989). Categorical perception of rhythmic patterns. *Psychological Research, 51*(1), 10-15.

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.

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.

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.

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.

Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology, 36*(1), 1-2.

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.

Varma, S., & Karl, S. R. (2013). Understanding decimal proportions: Discrete representations, parallel access, and privileged processing of zero. *Cognitive Psychology, 66*(3), 283-301.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "p< 0.05". *The American Statistician, 73*(S1), 1-19.