A "$p < .05$" Boundary Effect in the Encoding and Retrieval of p-values from Scientific Texts

V.N. Vimal Rao, raovnv@illinois.edu
Department of Statistics, University of Illinois, Urbana Champaign
Urbana, IL 61801 USA

Jeffrey K. Bye, jkbye@csudh.edu
Department of Psychology, California State University, Dominguez Hills
Carson, CA 90747 USA

Abstract

'Statistical significance' is more than just a label. Cognitive psychological theories suggest it may represent a mental concept of a category of p-values due to the pervasive practice of dichotomous interpretation of p-values. This paper builds on previous research identifying categorical boundary effects in the initial information processing of p-values by examining the encoding and retrieval of p-values embedded in the context of scientific abstracts. A sample of 30 U.S. graduate students in the psychological sciences read blocks of abstracts, then were prompted to recall certain details, including p-values. Results show that memory for p-values was skewed away from the .05 boundary, suggesting that training in dichotomous '$p < .05$' thinking may lead to categorical biases in memory for p-values. These results set up experiments to test mechanistic hypotheses of boundary effects on statistical cognition as well as the efficacy of teaching interventions to address and ameliorate these categorical biases.

Keywords: statistical thinking; mathematical cognition; concepts and categories; memory errors

Introduction

'Statistically significant' — two words that were always meant to be a simple guidepost in the interpretation of p-values as probabilities — have, since their introduction in the late 1800s, taken on a life of their own. Having sparked controversies (e.g., Hales & Wood, 2022) and questionable research practices (see John et al., 2012), scholars agree something must be done to change the current practice of statistical testing based on a dichotomous '$p < .05$' boundary.

Suggested solutions vary greatly. Cumming (2014) suggests an emphasis on confidence intervals and effect sizes. Kruschke and Liddell (2018) suggest a transition to the Bayesian school of statistics. Some statisticians advocate other statistics, such as second-generation p-values (SGPV; Blume et al., 2019), Shannon information (S-value; Greenland, 2019), and the false positive risk (FPR; Colquhoun, 2019).

All these suggestions, without proper framing, may fall prey to the same dichotomization problem as p-values – all are artificial measures utilizing scales that require helpful guideposts to support scientists' decision making, which in turn makes them susceptible to categorization by novices to facilitate interpretation. Lost in the midst is the core principle of statistical thinking – reasoning about uncertainty.

What these recommendations fail to consider is what if any permanent damage has been done by the previous century of life in the '$p < .05$' world. From a cognitive science perspective, there is reason to suspect that two innocuous words, 'statistically significant', have transcended to a category label impacting our cognitive processes. Indeed, words affect the way we think and process information (Hussein, 2012), and cognition may even fundamentally be an act of categorization (Harnad, 2017).

If the past century of statistical practice has created a mental representation of 'statistically significant' p-values and those that are not, simply ceasing to use the term will not suffice in reversing the psychological effects of '$p < .05$' dichotomous thinking. Furthermore, training efforts would require a cognitive component, to ensure that categorical boundary effects inherent to cognition can be developed in a manner consistent with recommendations for statistical practice.

The purpose of this paper is thus to explore the potential cognitive effects of having lived, even if only a few years, in a '$p < .05$' world. Specifically, we examine the potential effect of the 'statistically significant' category on early career researchers' encoding and retrieval of p-values embedded within scientific abstracts. These results can inform whether further education and self-regulatory strategies alone will suffice in moving to a world beyond '$p < .05$', or if additional cognitive training activities may be needed.

Background

Wasserstein et al. (2019), in addition to calling for an end to the era of 'statistical significance', give a warning for the future – "to preclude a reappearance of this problem elsewhere, we must not begin arbitrarily categorizing other statistical measures" (p. 2). However, to many cognitive scientists, cognition is categorization – perception of any stimulus, and in general the act of seeing a thing as 'something', is at its heart an act of categorization (Goldstone et al., 2013). Categorization is fundamental to our interaction with any attribute to which we differentially respond (Harnad, 1987), such as the interpretation of a p-value (or effect size, Bayes Factor, etc.) based on its magnitude.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Categorical Thinking

Categories provide structure to our interactions with the external world, facilitating the mapping of stimuli to responses by making connections between stimuli that may generate similar responses for a particular purpose. In this way, categorization can make different stimuli psychologically alike (Sloman, 1996). Some argue that the creation of a label initiates the creation of a mentally possessed notion, inviting categorization by creating both category and concept (Waxman & Markow, 1995). The existence of category labels then provides a psychological anchor for the subsequent development of a concept, as objects are determined to either be a member of the category or not (Clark, 1997). Providing redundant labels serves to reinforce concepts (Lupyan et al., 2007) – word labels accompanied by other symbolic or auditory labels lead to faster concept acquisition and increased robustness of the concept. By contrast, unlabeled stimuli have less influence on how people categorize than those reinforce with labels (McDonnell et al., 2012). This collectively suggests that labels indeed matter.

The implications of a century of dichotomizing *p*-values are profound. Consider the modern graduate student. Even with the shifting landscape of statistical practice in the wake of the replication crisis, the fact remains that $p < .05$ is still widely taught even if for historical purposes, and the extant scientific literature of the past century is riddled with *p*-value categorization. Future students reading past literature will undoubtedly encounter the term 'statistically significant'. Seeing *p*-values described as 'statistically significant' may still initiate the formation of a category and concept based on this label; in past literature, such *p*-values were typically not even reported exactly but rather simply as "$p < .05$" or "$p > .05$". Similarly, seeing *p*-values represented by asterisks in a table, such as "*" for "$p < .05$" will reinforce these students' concept of 'statistically significant'. Furthermore, solely seeing (modern) papers without the label may not be enough to undo the formation of the concept.

Benchmarks and Boundaries

Simply refraining from specifying category-delineating boundaries (e.g., '$p < .05$') in favor of category benchmarks (e.g., $d = .5$ is a 'medium' effect) will not obviate the problem. Collins & Watt (2021) have already found evidence of categorical effects in the interpretation of effect sizes based on Cohen's (1988) suggested values of .2 for a small effect, .5 for medium, and .8 for large[1].

Even in the absence of pre-specified benchmarks, individuals build a notion of category typicality through repeated exposure; whether by accrual of exemplars (Nosofsky, 1986) or representing a prototype (Rosch, 1975), individuals learn to categorize novel stimuli based on their experience. Furthermore, even when boundaries are not explicitly provided, stimuli at or near boundaries are often as effectively categorized as prototypes (e.g., Davis & Love, 2010). This may be because individuals automatically generate reference points for the boundary in order to help them discriminate between categories (Pastore, 1987).

Categorical Effects on Numerical Cognition

As statistics is fundamentally based on the utilization of numbers to quantify probability and uncertainty, statistical cognition is thus fundamentally based on numerical cognition. The classical model for the mental representation of numeric stimuli is a logarithmically compressed mental number line (Log MNL; Dehaene et al., 1990; Moyer & Landauer, 1967). First established with whole numbers, the Log MNL has since been extended to include decimal representations of numbers (e.g., Dehaene, 1997; Varma & Karl, 2013).

Specifically relevant for statistical cognition is the theory that the Log MNL is not a continuous representation, but rather, one with psychological boundaries based on our place-value system. This is most readily observed by the decade-crossing effect (Nuerk et al., 2015), the phenomenon by which determining the midpoint between two numbers is harder when the tens digits differ (i.e., bisecting '27 and 35' is harder than bisecting '21 and 29'). Similarly, children and adults are faster to indicate which two-digit number is larger when both the decade and unit digits are compatible (e.g., 37 > 25) than incompatible (e.g., 35 > 27), further suggesting that multi-digit numbers are not purely represented as holistic magnitudes (Nuerk et al., 2001).

To date, three studies have provided empirical credence to the potential of categorical distortions of the Log MNL. Landy et al. (2017) found evidence of a boundary effect between 999,999 and 1,000,000, conjecturing that the cause is likely to be the shift from the 'thousand' to 'million' category labels, and these labels' inciting of categorical effects on the Log MNL.

While prior research has identified a discontinuity around $p = .05$ in researchers' confidence in an effect (e.g., Rosenthal & Gaito, 1963), we have recently embarked on a series of lower-level examinations of researchers' initial processing of the *p*-values themselves. In two previous studies, we examined the effect of 'statistically significant' as a category on the initial processing of *p*-values (i.e., within the first seconds after exposure to the stimulus; Rao et al., 2022; 2024). We found evidence of a boundary effect at .05 in graduate students in the psychological sciences, even after adjusting for general landmark boundary effects inherent to the Log MNL; undergraduates only showed a tiny such effect, presumably based on the natural boundary of 5 in a base-10 system. Our conjecture was that the cause of this effect is training and exposure to the 'statistically significant'

[1] It should be noted that Cohen (1988) explicitly warned against adopting these values as benchmarks, as they were purely hypothetical, and should vary by field.

category and its '$p < .05$' boundary, and this category's inciting of categorical effects on the Log MNL.

Taken together, and in combination with the psychological theories of categories as well as numerical cognition, the empirical evidence from these studies supports the development of the following theory:

1. Exposure and training in the '$p < .05$' world incites a mental concept of 'statistical significance'.
2. This mental concept produces a psychologically real boundary at .05 delineating different categories of *p*-values, i.e., those that are 'statistically significant' and those that are not.
3. This psychologically real boundary governs individuals' interaction with *p*-values as numerical stimuli.

What remains to be seen is how much and over what time course such boundary effects occur for interpreting *p*-values. Here, we set out to test whether such effects occur in people's encoding and retrieval for specific *p*-values in scientific texts.

Problem Statement

Despite the ubiquity of categorization and the prevalence of the dichotomous interpretation of *p*-values in practice, it is important to note that statisticians and researchers do not, and (for the most part) never have, recommend the blind use of categories to interpret statistical measures – a cursory search of the literature will find many papers decrying such practices (e.g., Shaver, 1993). Why then do we still categorize statistical measures?

Based on psychological theories of categories and concepts, there are three key reasons why the categorization of statistical measures may be psychologically unavoidable. First, as some cognitive scientists have argued, categorization, including that based on numeric magnitude, may be innate to human psychology (Harnad, 2017). Alternatively, categorization may be a direct consequence of exposure to a past literature base that has extensively utilized categorical boundaries and benchmarks to guide the interpretation of statistical measures, such as '$p < .05$', 'RMSEA > .10', or '$d = .20$'. And finally, categorization may be a direct consequence of instruction and training that (by necessity) helps students learn how to differentially respond to statistical measures based on their numeric magnitude.

These accounts provide us with potential insights into how we move to a world beyond statistical dichotomization. If the categorization of statistical measures is inescapable, the goal should not be to avoid categorization, but rather to avoid letting categorization become a problem. In order to avoid repeating past mistakes akin to that of the *p*-value controversy, we must further investigate statistical cognition to better understand how and when the categorization of statistical measures lead to suboptimal outcomes in statistical thinking.

Research Question

The process of science is not one on the order of milliseconds, but often is on the order of months and years. On this timescale, a categorical boundary effect at .05 in the initial processing of *p*-values as stimuli (Rao et al., 2022; 2024) is only problematic for statistical and scientific practice if that categorical boundary also exists on these longer time scales, or if it has downstream effects on cognitive processes that are on those longer timescales. Cognition on these longer time scales utilizes additional cognitive faculties, such as memory, than those used in initial information processing.

To address this critical gap, the current study investigates the following research question: after controlling for boundary effects in numerical cognition, do graduate students with statistical training show categorical boundary effects at .05 in the recall of the numeric magnitude of *p*-values that are embedded in scientific abstracts?

Methods

The key experimental stimuli were a series of scientific abstracts that contained exact *p*-values. The experiment employed a within-participants design in order to control for individual differences in encoding and retrieval strategy and working memory.

Participants

We recruited 30 graduate students from the psychological sciences from two large public research-oriented universities in the United States. Eligible participants were those who reported having at least one full year of experience with hypothesis tests and *p*-values, as well as at least one year reading scientific abstracts through coursework, research, or teaching. Consenting participants were paid $25 USD for completing the 60-min study[2].

Materials

Each stimulus presented to participants was in the form of a short scientific abstract. All abstract vignettes contained information about a fictional study and were written to control for several key covariates in encoding and retrieval of texts (e.g., text structure, experimental vs. correlational design, IVs and DVs, sample size, etc.). The key attribute of each stimulus was the numerical value of the *p*-value contained therein.

We created the *p*-value stimuli to embed in scientific abstracts according to multiple criteria. First, due to the effect of leading zeros on numerical cognition (Schulze et al., 1991), all target *p*-value stimuli were between .010 and .099. Numeric values with repeated digits were excluded from consideration, as were values ending in 5, as such stimuli are easier to recall (Milikowski & Elshout, 1995). To reduce demand characteristics due to the similar range of target *p*-

[2] The study also included two additional tasks whose results are not reported here due to space limitations.

values, we also included four distractor stimuli with values below .010 or above .100.

Furthermore, to counter a possible 'rounding' strategy whereby participants may round each *p*-value to the nearest hundredths digit, or a possible 'hundredths digit' strategy whereby participants truncate and ignore the thousandths digit of the *p*-value, numerical values for the *p*-value stimuli were carefully selected to distinguish between predictions of a 'rounding' strategy, 'hundredths digit' strategy, and the 'categorical boundary' conjecture (see Table 1). Specifically, *p*-value pairs were presented to participants such that for a given pair, each *p*-value will be equidistant from .05, e.g., 'p = .027' and 'p = .073'. All participants were presented with the same *p*-value stimuli, but each *p*-value in a pair was counterbalanced to different abstract across participants.

Table 1: Numeric values of *p*-value stimuli.

Stimuli Group	Numeric Values
Pair #1	.014, .086
Pair #2	.021, .079
Pair #3	.027, .073
Pair #4	.036, .064
Pair #5	.039, .061
Pair #6	.042, .058
Distractors	.003, .172, .491, .528

The manner in which these stimuli controlled for the three possible recall strategies is as follows. For example, if participants employed a 'rounding' strategy, their category bias for Pair #1 would be -.008 (i.e., [.010 - .014] - [.090 - .086]). However, their category bias for Pair #3 would be .008 (i.e., [.040 - .036] - [.060 - .064]). In this manner, their average category bias would be 0. As a secondary protection against a 'rounding' strategy leading to a non-zero sample statistic, three of the 'statistically significant' *p*-value stimuli 'round up' while three 'round down', with the same balance in rounding for the 'not statistically significant' stimuli. Similarly, if participants employed a 'hundredths digit' strategy, their category bias for Pair #1 would be .002 (i.e., [.010 - .014] - [.080 - .086]), while their category bias for Pair #3 would be -.002 (i.e., [.030 - .036] - [.060 - .064]), again leading to an average category bias of 0.

Importantly, only through the conjectured 'category boundary' effect at .05 would participants have a non-zero average category bias across all 6 pairs. If the .05 boundary does impact participants' recall, we would expect a *positive* average category bias for 'not statistically significant' *p*-values, akin to the .05 boundary 'pushing' participants' recall *away* from .05 (e.g., .052 may be falsely recalled as .058). Similarly, we would expect a *negative* average category bias for the 'statistically significant' *p*-values, akin to the .05 boundary 'pushing' participants' recall *away* from .05 (e.g., .048 may be falsely recalled as .042). If this were true, participants' category bias for Pair #1 would be negative, perhaps -.008 as in the previous example (i.e., [.010 - .014] - [.090 - .086]), and their category bias for Pair #3 would also be negative, perhaps -.008 (i.e., [.032 - .036] - [.068 - .064]), uniquely leading to a negative average category bias.

As all *p*-value stimuli were embedded in text in the form of a scientific abstract, we also created the texts for each stimulus according to multiple criteria. In order to control for variation in text comprehension, all stimuli followed the same text structure (see Figures 1 and 2). All stimuli utilized contexts that were designed to be accessible and easily understandable for all study participants.

> Does yoga reduce 5k running time? A recent study recruited 84 adults and randomly assigned them to either a 20-minute yoga group or a control group. Participants then completed a 5k race. Relative to the control group, participants in the yoga group had statistically significantly lower 5k run times (p = .039). Results suggest that yoga may decrease 5k running time.

Figure 1: Example experimental study abstract stimulus.

> Is students' college GPA related to their graduate school GPA? A recent study recruited 597 students from across the US. Researchers collected their GPA at the time of graduation from college and grad school. Students' college GPA was not statistically significantly correlated to their grad school GPA (p = .064). Results suggest that students' college GPA may not be related to grad school GPA.

Figure 2: Example observational study abstract stimulus.

Each abstract began with a research question, in interrogative form, listing the two factors being studied. Half of these questions implied an experimental study by framing the question causally (e.g., "reduce", Figure 1), while the other half implied an observational study by framing the question correlationally (e.g., "related to", Figure 2). Next, the abstract provided information about the sample size and study method, which reinforced either the experimental framing (e.g., "randomly assigned", "control group") or observational (e.g., "collected"). The abstracts then presented the results of the statistical analysis in the form of a *p*-value. Finally, the abstracts ended by drawing a conclusion based on the statistical analysis. Importantly, no special attention was drawn to any specific element of the abstract. As prior knowledge and level of interest in the contexts utilized may also affect the encoding and retrieval of p-values, we isolated the effect of context by counterbalancing the *p*-value and context pairings when presenting stimuli to participants.

Procedure

Participants were given a single practice stimulus to familiarize themselves with the format of the study. The practice stimulus contained an excerpt from a poem by Maya Angelou, and participants were subsequently asked three short questions about the excerpt. After completing this practice stimulus, participants were then presented the 16

abstract stimuli in four blocks of four abstracts each. Stimuli were a priori assigned to a block based on the stimuli characteristics (e.g., research question, *p*-value, study design), but the order of stimuli within each block was randomly presented to participants. Participants were instructed to "Remember to pay attention when reading each abstract in order to answer questions about them later in the study". Again, no special attention was called to the *p*-values relative to the other features of the abstracts.

Within each block, participants were shown each abstract one at a time and were allowed to read each abstract for as long as they desired. After the text of the fourth abstract in the block, participants completed a Flanker Task (Wilhelm et al., 2013), responding to 20 stimuli each time they completed the task. The purpose of completing the Flanker Task was to tax participants' working memory, minimizing rehearsal of the information contained in the abstract stimuli, ensuring that their answers to the recall questions elicited encoding and retrieval cognitive processes. A non-verbal, non-numerical distractor task was used to minimize potential carryover effects from the distractor task to their memory for the abstracts.

After completing the Flanker Task, participants were then asked six questions about each of the four abstracts in the block (see Table 2), in the order in which the abstracts were originally presented. The participants were instructed to "Provide your best guess when answering the questions" and "Only if you truly do not remember at all, please write 'don't remember'." Additionally, after completing the recall questions for the second block, participants were also told that "You may now take a short break of 2-3 minutes". This instruction was provided in order to mitigate any working memory fatigue participants may have experienced.

Table 2: List of recall questions.

Recall questions
What were the variables being considered?
What was the study design?
Who were the participants?
What was the sample size?
What was the result/finding of the study?
What was the *p*-value?
Please write anything else you remember about the study:

Note. As in the abstract stimulus presentation, no special attention was called to the *p*-value at recall.

Power Analysis

Participants' recall bias was calculated as the numeric value the participants recalled *minus* the actual numeric value of the stimulus. In a study in which each participant sees 12 experimental stimuli (i.e., six *p*-value pairs), in order to detect a medium effect quantified as an average categorical bias of .001 against random errors distributed normally with mean 0 and standard deviation of .002 in the between-participant paired difference in average categorical bias with 90% power and 90% accuracy, at least 32 participants would need to be recruited. Participant recruitment was temporarily stopped at $n = 30$ due to payment funds lapsing at the end of the year; a follow-up study is currently planned.

Results

After adjusting for other potential effects, participants' recall of the numerical magnitude of the *p*-value stimuli were biased away from the .05 boundary (see Figure 3 and Figure 4). 'Statistically significant' *p*-values were on average recalled as a lower numerical magnitude than presented (average percent error in recall = -4.6%), while 'not statistically significant' *p*-values were on average recalled as a slightly higher numerical magnitude (average percent error = 0.3%). This is unlikely due to participants rounding to the nearest hundredths digit, as on average, participants' recall of the stimuli p=.027, p = .036, and p = .039 were all 'rounded down', while their recall of the stimuli p=.061, p = .064, and p = .073 were all 'rounded up' (see Table 2).

Additionally, this effect is not consistent within each *p*-value category – the closer the numerical magnitude of a *p*-value is to the boundary, the larger the error in recall, with larger negative errors for the 'statistically significant' *p*-values just below .05, and larger positive errors for the 'not statistically significant' *p*-values just above .05 (see Figure 4). This is akin to a repelling effect of the .05 boundary, consistent with the cognitive theory of category boundaries.

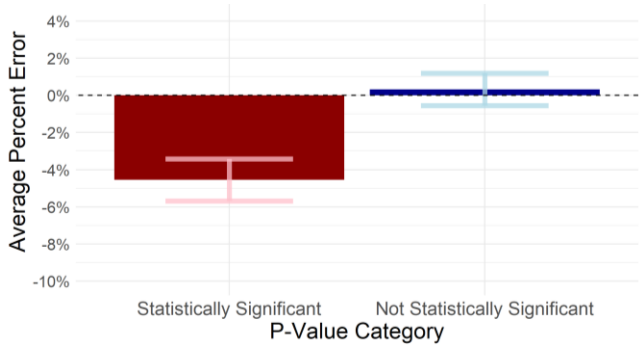

Figure 3: Participants' average recall by *p*-value category.

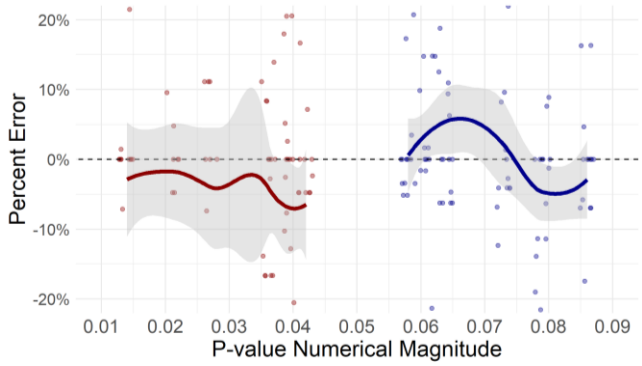

Figure 4: Participants' recall error by *p*-value magnitude.

Table 2: Average recall error by *p*-value stimuli.

Stimuli Pair	Avg. Percent Error
.014, .086	-3.2%, -2.3%
.021, .079	0.7%, -7.1%
.027, .073	-4.0%, 2.0%
.036, .064	-3.9%, 6.0%
.039, .061	-6.8%, 2.5%
.042, .058	-6.5%, 1.0%

Interestingly, the pattern appears to hold true for those *p*-value pairs closest to the .05 boundary, but not those pairs further away – this is also consistent with the theory of benchmarks and boundaries in which the boundary facilitates efficient categorization, and all members of categories are 'pushed' or 'pulled' towards benchmark values within each category.

Finally, we fit a linear mixed effects model with the percent error in recall as the response variable, the *p*-value category and the numerical magnitude of the *p*-value as a fixed effects, and the participant ID and the stimulus pairings as random effects. Results from the model suggest that there is indeed on average a categorical boundary effect in the retrieval of the numerical value of *p*-values. Per the model results, 'Statistically significant' *p*-values were more likely to be recalled as a lower numerical magnitude than 'not statistically significant' *p*-values by a difference of approximately -13.6% error (95% CI: -22.0% – -5.1%; p = .0019).

Discussion

This study expands on our previous work (Rao et al., 2022; 2024) by examining whether the categorical boundary effect at p = .05 found in emerging scientists' initial information processing of *p*-values as numeric stimuli extends to their cognition on longer and more ecologically valid time-horizons. Specifically, this study attempted to identify the presence of a categorical boundary effect at p = .05 in the encoding and retrieval of the numeric magnitude of *p*-values embedded within scientific abstracts. In fact, such an effect was present.

The finding of a categorical boundary effect suggests that the category membership of a *p*-value as either 'statistically significant' or 'not statistically significant' exerts an effect on the way statistically trained graduate students remember *p*-values across a timescale of several minutes.

We hypothesize that this may be due to a discontinuity in graduate students' mental representation of the *p*-value continuum causes a distorted initial information processing of the numeric magnitude of *p*-values (as found in our previous studies) which subsequently is encoded and retrieved in a distorted manner. Additionally, it is possible that the '$p < .05$' heuristic, combined with the semantic category label 'statistically significant', is too pervasive for graduate students to override via self-regulation during the encoding process. These hypotheses can be tested in future studies. Specifically, if the latter hypothesis is true, graduate students' scores on a measure of inhibitory control (such as the Flanker Task) should be correlated with their average categorical bias in *p*-value recall. We will test this prediction in a study with a larger number of participants.

Additionally, if the former hypothesis is true, the categorical boundary effect in graduate students' initial information processing of *p*-values in a task such as those used in our previous studies should be correlated with their average categorical bias in *p*-value recall in the abstract task utilized in this study. We plan to assess this possibility in an upcoming follow-up to our current sample.

Furthermore, graduate students exposed to a categorical training activity (such as those conducted by Goldstone, 1994, but adapted for *p*-value stimuli) should show a reduced categorical boundary effect in their initial information processing of *p*-values, and should thus also show a reduced average categorical bias in *p*-value recall. We plan to test these predictions in future studies.

The current experiment has several limitations. Most significantly, it only included results from 30 participants who were only provided 12 experimental stimuli each. Thus, the patterns identified in the data cannot confidently be separated from expectations due to random variation and inherent individual differences. Additionally, this smaller-than-intended sample size inhibited our planned analysis. Future analyses will explore patterns with more sophisticated statistical models. Finally, absent a control group, it is difficult to determine the extent to which the identified effect is due to a categorical bias due to statistical training and practice in a '$p < .05$' world, against the extent to which it is a natural consequence of Log MNL on encoding and retrieval of numeric magnitudes in general, although this possibility is somewhat minimized by our prior research (Rao et al., 2024) in which we found that statistically-untrained undergraduates show only a minimal boundary effect around .05.

Summary and Conclusion

Recently, Wasserstein et al. (2019) called for an end to the era of 'statistical significance', urging that *p*-values always be presented as a numeric magnitude, rather than a priori categorized as '$p < .05$' or '$p < .01$'. This advice is meant to eschew dichotomous statistical thinking and decision making in favor of 'Accepting uncertainty', 'being Thoughtful', 'being Open', and 'being Modest' – ATOM. However, categorizations are generally helpful to novices (Gibson, 1969), and once entrenched do not easily fade (McDonnell et al., 2012), although they can be mitigated with targeted training (e.g., Goldstone, 1994).

This study provides an initial investigation into the mental processes associated with reading and remembering *p*-values as part of scientific abstracts. The results show that categorical biases based on the '$p < .05$' boundary may exist for emerging psychological scientists even when *p*-values are presented as numeric magnitudes. Our results set the stage for future research on the cognitive processes underlying statistical thinking as well as educational interventions to support moving to a world beyond '$p < .05$'.

Acknowledgments

We thank Ali Fulsher, Rina Harsch, and Mónica González-Márquez for helpful feedback on the scientific abstract materials and recall questions.

References

Blume, J. D., Greevy, R., A., Welty, V. F., Smith, J. R., & Dupont, W. D. (2019). An introduction to second-generation p-values. *The American Statistician*, *73*(S1), 157-167. https://doi.org/10.1080/00031305.2018.1537893

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Collins, E., & Watt, R. (2021). *Use, knowledge, and misconceptions of effect sizes in psychology*. Preprint from PsyArXiv. https://doi.org/10.31234/osf.io/r7vmf

Colquhoun, D. (2019). The false positive risk: A proposal concerning what to do about p-values. *The American Statistician*, *73*(S1), 192-201. https://doi.org/10.1080/00031305.2018.1529622

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

Davis, T., & Love, B. C. (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science*, *21*(2), 234–242. https://doi.org/10.1177/0956797609357712

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. New York: Oxford University Press.

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. https://psycnet.apa.org/doi/10.1037/0096-1523.16.3.626

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://psycnet.apa.org/doi/10.1037/0096-3445.123.2.178

Goldstone, R. L., Kersten, A., & Carvalho, P. F. (2013). Concepts and categorization. In A. F. Healy, R. W. Proctor, & I. B. Weiner (Eds.), *Handbook of psychology: Experimental psychology* (pp. 607–630). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119170174.epcn308

Greenland, S. (2019). Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with S-values. *The American Statistician*, *73*(S1), 157-167. https://doi.org/10.1080/00031305.2018.1529625

Hales, A. H., & Wood, N. R. (2022). Statistical Controversies in Psychological Science. In *Avoiding Questionable Research Practices in Applied Psychology* (pp. 191-211). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-04968-2_9

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

Hussein, B. A. S. (2012). The sapir-whorf hypothesis today. *Theory and Practice in Language Studies*, *2*(3), 642-646. https://doi.org/10.4304/tpls.2.3.642-646

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*(5), 524-532. https://doi.org/10.1177/0956797611430953

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic bulletin & review*, *25*, 178-206. https://doi.org/10.3758/s13423-016-1221-4

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

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is Not Just for Talking: Redundant Labels Facilitate Learning of Novel Categories. *Psychological Science*, *18*(12), 1077-1083. https://doi.org/10.1111/j.1467-9280.2007.02028.x

McDonnell, J. V., Jew, C. A., & Gureckis, T. M. (2012). Sparse category labels obstruct generalization of category membership. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*, 749-754. Retrieved from https://escholarship.org/uc/item/5wj0m4g2

Milikowski, M., & Elshout, J. J. (1995). What makes a number easy to remember?. *British Journal of Psychology*, *86*(4), 537-547. https://doi.org/10.1111/j.2044-8295.1995.tb02571.x

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

Nosofsky, R. M. (1986). Attention, Similarity, and the Identification–Categorization Relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nuerk, H. C., Moeller, K., Klein, E., Willmes, K., & Fischer, M. H. (2015). 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

Pastore, R. E. (1987). Categorical Perception: Some Psychophysical Models. In Harnad, S. (Ed.) *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

Rao, V. N. V., Bye, J. K., & Varma, S. (2022). Categorical Perception of p-Values. *Topics in cognitive science*, *14*(2), 414-425. https://doi.org/10.1111/tops.12589

Rao, V. N. V., Bye, J. K., & Varma, S. (2024). The psychological reality of the learned "p < .05" boundary. *Cognitive Research: Principles and Implications*, *9*(1), 27. https://doi.org/10.1186/s41235-024-00553-x

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233. https://doi.org/10.1037/0096-3445.104.3.192

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *The Journal of Psychology*, *55*, 33–38. https://doi.org/10.1080/00223980.1963.9916596

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *The Journal of Experimental Education*, *61*(4), 293-316. https://doi.org/10.1080/00220973.1993.10806592

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

Sloman, S. A. (1996). The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

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

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. https://doi.org/10.1080/00031305.2019.1583913

Waxman, S. R., & Markow, D. B. (1995). Words as Invitations to Form Categories: Evidence from 12- to 13-month-old Infants. *Cognitive Psychology*, *29*(3), 257–302. https://doi.org/10.1006/cogp.1995.1016

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it?. *Frontiers in psychology*, *4*, 433. https://doi.org/10.3389/fpsyg.2013.00433