GRADUATE STUDENTS' EFFECT SIZE CATEGORY BOUNDARIES

V.N. Vimal Rao[1], Jeffrey K. Bye[1], and Sashank Varma[2]
[1]Department of Educational Psychology, University of Minnesota
[2]School of Interactive Computing & School of Psychology, Georgia Institute of Technology
rao00013@umn.edu

Statisticians increasingly decry ritualistic categorizations of statistical measures. The interpretation of effect sizes is often guided by benchmarks, i.e., d = .2 ('small'), .5 ('medium'), and .8 ('large'). We employed a cognitive psychology approach to investigate how researchers systematically categorize values between these benchmarks. We find effect size categories are separated by fuzzy boundaries – as predicted by psychological theories of categorization. Understanding the cognitive processes underlying statistical reasoning can help us consider how to move beyond ritualistic interpretation of statistical measures.

INTRODUCTION

In 2019, Wasserstein, Schirm, and Lazar, on behalf of the American Statistical Association, called for an end to the era of statistical significance. As many fields have moved to emphasize effect sizes (e.g., Cumming, 2014), Wasserstein et al. additionally give a warning for the future – "to preclude a reappearance of this problem elsewhere, we must not begin arbitrarily categorizing other statistical measures" (Wasserstein et al., 2019, p. 2).

However, to cognitive psychologists, categorization is fundamental to cognition – perception of a stimulus and seeing it *as something* is, at its heart, an act of categorization (Goldstone et al., 2013). Categorization naturally emerges whenever we respond differently to objects based on some attribute (Harnad, 1987), such as the interpretation of an effect size based on its numerical magnitude.

In this paper we first consider why individuals might inherently categorize statistical measures. We then examine whether the widespread use of benchmarks underlies categorical interpretations of effect sizes in a manner antithetical to Wasserstein et al.'s (2019) warning against categorization.

BACKGROUND

Categorization is ubiquitous to cognition. How might it present itself during acts of statistical thinking? Categories are collections of objects in the world. Concepts are mental representations of these categories: They denote what objects are being represented and how that information can be used to make inferences (Smith, 1989). Concepts provide structure to our interactions with the external world:

- Concepts efficiently encode information, reducing cognitive processing (e.g., Bruner et al., 1956; Goldstone et al., 2013) – rather than storing complete information about every right skew distribution one has encountered, one need only store a single representation of a *right skew* distribution.
- Concepts facilitate the generalization of experiences to objects within the same category (e.g., Goodman et al., 2008) – the concept of *multicollinearity* provides information regarding the interaction between two explanatory variables in a statistical model.
- Individuals who share concepts can succinctly communicate information with one another (e.g., Markman & Makin, 1998) – describing a variable as a *confounder* to a fellow statistician (who shares the concept of a confounder) communicates information about its relationship with other variables.

Cognitive psychologists generally accept the ubiquity of concepts and view a wide variety of cognitive acts as fundamentally an act of categorization (Murphy, 2002).

Benchmarks and Boundaries

Benchmarks serve as the most typical object belonging to a category, i.e., they are the *prototype* for the concept. Benchmarks can either be explicitly specified (e.g., .5 is the prototypical 'medium' effect size) or formed implicitly, for example as the weighted average of all members of the category (Rosch, 1975). Once these benchmarks are identified, individuals use them to determine category membership of novel stimuli based on the similarly of these stimuli to the benchmark (Rosch & Mervis, 1975) – a 'fuzzy' boundary.

For example, when determining whether a given distribution is normal, statisticians may compare it to the mathematically defined normal distribution, which serves as the prototypical normal distribution. If a given distribution is similar enough to this benchmark, we can treat the distribution in the same way we would treat the prototypical normal distribution.

Even in the absence of pre-specified benchmarks, individuals build a notion of category typicality through repeated exposure (Posner & Keele, 1968). The degree to which a new stimulus is *similar* to a benchmark is based on the extent to which the behavioral responses are similar (Palmeri, 1997). For example, when examining the normality of residuals from a simple linear regression model, distinguishing when the distribution is 'acceptable' and when some remedial action must be taken helps form a cohesive concept of *normal* with which to categorize residual plots. Through such a repeated exposure to both stimuli and their associated responses, individuals are able to delineate distinct categories of stimuli and form benchmarks for concepts.

In contrast to the prototype model of categories, the category boundary model holds that a concept representation describes the boundary around a category, and therefore specifying a boundary leads to the formation of a concept (Ashby, 1992). For example, .05 is a common boundary delineating 'statistical significance', and while individuals may possess this concept, they may not possess a prototypical 'statistically significant' *p*-value. However, even when boundaries are not explicitly provided, stimuli at or near boundaries are sometimes categorized as effectively as prototypes (e.g., Davis & Love, 2010). In these cases, stimuli near a categorical boundary are treated similarly to its category's benchmark – a 'hard' boundary.

A New Statistics with Old Problems?

Much like the *p*-value controversy where 'statistical significance' created a publication bias against studies with large *p*-values, there is already evidence of a burgeoning effect size controversy replete with its own publication bias. For example, Schäfer and Schwarz (2019) found a problematic difference in the distribution of effect sizes between publications with pre-registration and those without. Is this because individuals are already categorizing effect sizes, like they categorized *p*-values? Consistent with this possibility, Collins and Watt (2021) found that the overwhelming majority of psychology researchers they surveyed consider the values provided by Cohen (1988) as best exemplifying 'small', 'medium', and 'large' effect sizes, despite Cohen's warning that these values were arbitrarily chosen.

It is currently unknown whether individuals consistently draw boundaries between these effect-size categories, and if so, where. Psychological research on categorization suggests that the existence of common benchmarks will lead individuals to delineate between concepts with boundaries. This is especially true for students, as novices tend to categorize stimuli based on superficial features, and these superficial features produce distinctive concepts (Chi et al., 1981).

The reification of categories and concepts in the interpretation of statistical measures can alter the manner in which individuals perceive stimuli. This phenomenon, called a categorical perception effect, results in exaggerated perceived differences across categories and diminished perceived differences within a category. These effects have been documented in the initial processing of *p*-values (Rao et al., 2022).

It is possible that through repeated instruction and practice, the widespread familiarity with Cohen's *d* benchmarks may reinforce the cognitive concept of 'small', 'medium', and 'large' effect sizes, and this may in turn induce a categorical perception effect in the interpretation of effect sizes. However, unlike *p*-values, where a clear boundary is provided, Cohen's *d* effect size categories are defined by benchmarks, i.e., d = .2 ('small'), d = .5 ('medium'), and d = .8 ('large'). These benchmarks are widely known (Collins & Watt, 2021), although it is unknown how researchers systematically categorize values falling between these effect size benchmarks.

Therefore, the purpose of this study is to examine where and how researchers draw boundaries between effect size categories: at what magnitude does an effect size 'change' from being categorized as 'small' to 'medium' and from 'medium' to 'large', and is this change gradual (as with a fuzzy boundary) or immediate (as with a hard boundary)?

METHODS

To identify the location of boundaries between effect size categories, we employed a cognitive psychology approach. Boundary identification tasks are commonly used in the study of categorical perception effects. They are often the first step in evaluating the cognitive effects of categories and concepts on individuals' interactions with stimuli (e.g., what hue(s) form(s) the boundary between blue and green?).

Graduate students in the psychological sciences at a research university in the Midwestern United States were recruited for this study (*n* = 39). All participants had completed at least one year of instruction and training in statistical methods at the doctoral level. They completed the boundary identification task as the second of three tasks. The full study took approximately 40 minutes on average to complete in full, and participants were compensated with a $25 electronic gift card.

Participants were told that they will be shown "values of Cohen's *d*, a statistic indicating the size of an effect in standard deviation units". They were then told that for each value, they would judge whether it indicated no effect, a small effect, a medium effect, or a large effect, by selecting one of four keys on a keyboard. Crucially, participants were not told how to make this judgment, and at no point in the study were the standard benchmarks (i.e., .2, .5, and .8) mentioned to participants.

Participants completed 180 trials in four blocks of 45 stimuli each, preceded by eight practice trials. There were 90 unique stimuli of the form "*d* = .XX", with values ranging from .01 to .90. Participants categorized each value twice: once in the first two blocks and again in the last two. Within each set of two blocks, the stimuli order was randomly shuffled. Stimuli were presented one-at-a-time and remained on screen until participants made their selection. Participants were encouraged to make their initial selection as quickly as possible.

After completing the full study, participants also completed a short survey collecting basic demographic information and probing for possible demand characteristics for the study. As part of this survey participants were asked to specify the upper and lower bounds of what they would consider a 'small', 'medium', and 'large' Cohen's *d* effect size.

RESULTS

To identify the point at which participants' effect size categories 'changed' from being categorized as 'small' to 'medium' and from 'medium' to 'large', we first analyzed their survey responses. Of the 39 participants, 34 self-reported that they referred to the values of .2, .5, and .8 when categorizing effect sizes. The remaining five participants referred to the values .1, .3, and .5 – common benchmarks in the interpretation of correlation coefficients (i.e., *r*). Data from these five participants was analyzed separately.

Participants' self-reported categorical boundaries were varied, with the median boundary delineating 'small' and 'medium' effect sizes at .39, and the median boundary delineating 'medium' and 'large' effect sizes at .70. Surprisingly, very few participants specified categorical boundaries at the midpoint between common benchmark values (i.e., .35 and .65; see the left panel of Figure 1). This may reflect a variety of interpretations of the benchmark values of .2, .5, and .8. Participants drawing a boundary between 'small' and 'medium' effect sizes near .5 might have interpreted .5 as a boundary rather than a benchmark, as is typical of other statistical measures such as *p*-values (where .05 serves as a categorical boundary). Participants drawing a boundary between 'small' and 'medium' effect sizes near .3 might be due to the desirability of finding a 'medium' effect rather than a 'small' one. Participants drawing a boundary between 'small' and 'medium' effect sizes near .4 might be due to a conservative approach based on an aversion to taking the risk of over-interpreting statistical results and possibly committing a questionable research practice.

Participants' estimates on the survey, on the aggregate level, matched their performance on the boundary identification task. As seen in the right panel of Figure 1, participants' responses showed average categorical boundaries at .38 (delineating 'small' and 'medium' effect sizes) and .69 (delineating 'medium' and 'large' effect sizes), consistent with the average self-reported values. Interestingly, there is quite a bit of overlap in the assigned labels for a given effect size. This may either be due to the psychological boundary between effect size categories indeed being a fuzzy boundary, or due to variability in the location of hard boundaries amongst participants, as observed in the self-reported boundary values.

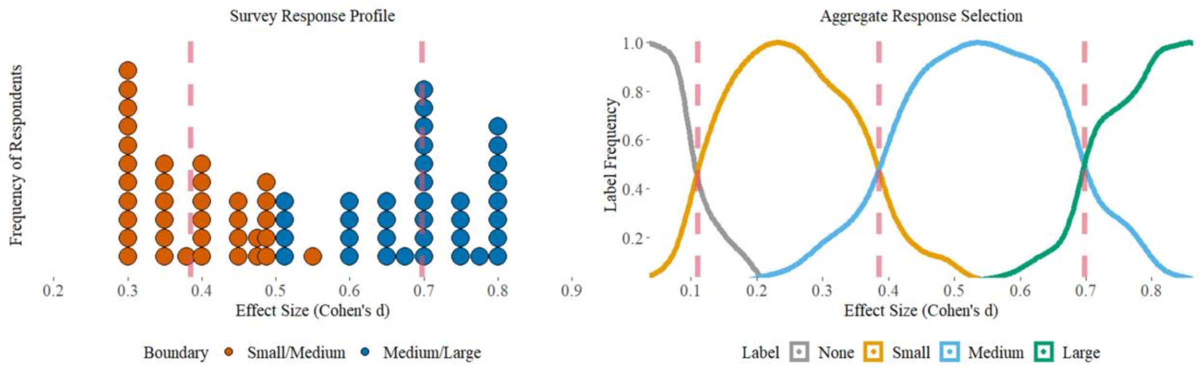

Figure 1: *Aggregate response patterns – Boundary location survey responses with reference lines at aggregate boundaries (left) and boundary identification task responses with reference lines at aggregate boundaries (right)*

Participants categorized each effect size between .01 and .90 twice, but their two responses were not always in agreement (see the left panel of Figure 2). The average agreement rate across all participants and effect size values was approximately 88%. The agreement rate was lower (as low as 60%) near boundaries (i.e., values of .11, .38, and .69, as identified in the aggregate response selections), and higher near benchmarks (i.e., values of .2, .5, and .8). Interestingly, the location of the *most* consistently categorized effect sizes were not the benchmark values themselves, but rather at slightly higher values.

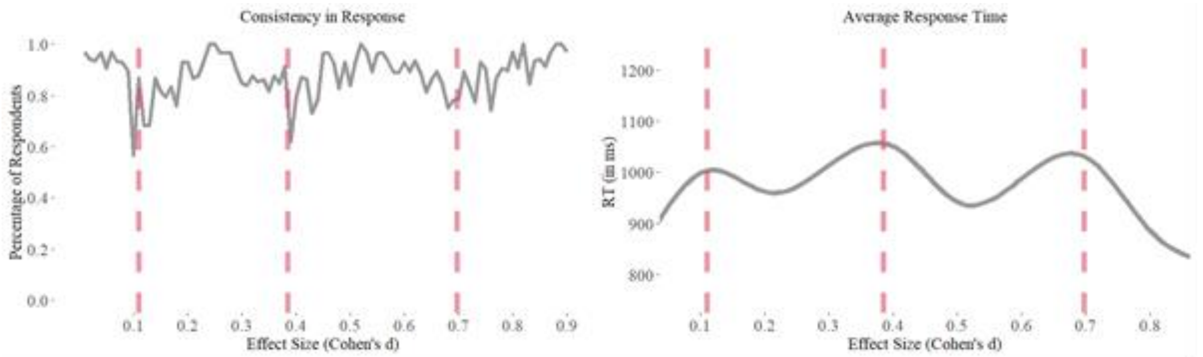

Figure 2: *Response patterns – Consistency in response selections and average response time by effect size value with reference lines at aggregate boundaries*

This pattern is also seen in participants' response times (see the right panel of Figure 2). Participants' response times in selecting a category label were approximately 12% slower when categorizing values near boundaries, relative to benchmark values. However, participants were fastest in making their selection at values slightly higher than the benchmark values. These response patterns exhibit boundary effects consistent with a 'fuzzy' boundary model.

While aggregate results indicate that participants' psychological boundaries are fuzzy, an examination of individual participants' response profiles indicates that for at least some participants, the boundaries are hard boundaries. The hard boundary was most commonly observed for the 11 participants who interpreted common benchmark values as boundaries, rather than drawing boundaries between benchmark values. For example, the response profile of participant GID1 (see the top left panel of Figure 3) indicates a hard boundary between 'medium' and 'large' Cohen's *d* effect sizes at .5. Similarly, .5 serves as a hard boundary between 'small' and 'medium' Cohen's *d* effect sizes for participant GID38 (see the top right panel of Figure 3). However, not all categorical boundaries were hard for these participants, as the boundary between 'small' and 'medium' effect sizes appears to be a fuzzy boundary for GID1, evidence by the overlap in the category labels assigned to each effect size. Similarly, the boundary between 'medium' and 'large' effect sizes is a fuzzy boundary for GID38.

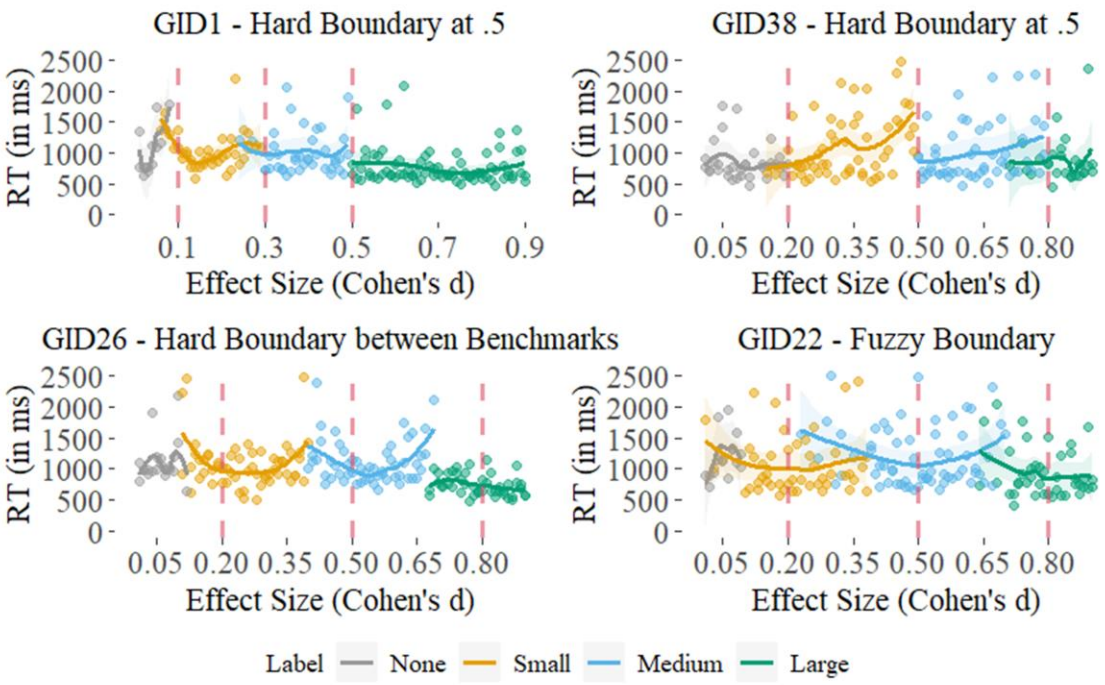

Figure 3: *Selected participant response patterns – Participants' responses (label) and response time (RT) with reference lines at common benchmark values as provided by Cohen (1988)*

Some participants who interpreted the values of .2, .5, and .8 as benchmarks still drew hard boundaries between categories. For example, participant GID26's response profile (see bottom left panel of Figure 3) shows the benchmarks relatively centered within each category, and hard boundaries between categories at .11, .40, and .68 respectively. However, GID26's response times were generally longer for effect size values near their category boundaries than for values near category benchmarks, consistent with a 'fuzzy' boundary model of concepts.

Most participants' (23 of 39) response patterns clearly reflected a fuzzy boundary between categories, as exemplified by the response profile of GID22 (see bottom right panel of Figure 3). These response profiles exhibit overlap between categories as well as increased response times and decreased consistency near category boundaries.

DISCUSSION

In this study we investigated how researchers categorize effect sizes into the commonly utilized 'small', 'medium', and 'large' categories. We find effect size categories are typically (but not always) separated by 'fuzzy' boundaries – as predicted by psychological theories of categorization.

Surprisingly, participants' response patterns and survey responses indicate that participants do not draw boundaries *exactly* at the arithmetic midpoints between common benchmark values, nor are they fastest and most accurate at *exactly* the common benchmark values. This may be due to the way in which we perceive symbolic (and non-symbolic) numbers – the standard model of numerical cognition suggests we possess a logarithmically compressed mental number line with psychological boundaries based on our place value system (Moyer & Landauer, 1967; Nuerk et al., 2011; Varma & Karl, 2013). Therefore, participants' boundaries may reflect psychological midpoints based on their mental number line.

This study is the first to empirically explore how researchers categorize a wide range of effect sizes, specifically in how they draw boundaries between effect size categories. Using labels such as those commonly utilized for Cohen's *d* effect sizes affects not only students' benchmarks but also the boundaries between them, sometimes in unpredicted ways. Understanding the cognitive processes underlying statistical reasoning can inform what we should practice and what we should teach if we are to move beyond the ritualistic categorical interpretation of statistical measures.

REFERENCES

Ashby, F. G. (1992). *Multidimensional models of perception and cognition*. Lawrence Erlbaum Associates, Inc.

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

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121-152. https://doi.org/10.1207/s15516709cog0502_2.

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.

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.

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.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154. https://doi.org/10.1080/03640210701802071.

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.

Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General*, *127*(4), 331–354. https://doi.org/10.1037/0096-3445.127.4.331.

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

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

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.

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(2), 324–354. https://doi.org/10.1037/0278-7393.23.2.324.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3, Pt.1), 353–363. https://doi.org/10.1037/h0025953

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.

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.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605. https://doi.org/10.1016/0010-0285(75)90024-9.

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in psychology*, *10*, 813. https://doi.org/10.3389/fpsyg.2019.00813.

Smith, E. E. (1989). Concepts and induction. In M. I. Posner (Ed.), *Foundations of Cognitive Science* (pp. 501–526). MIT Press.

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.