WRITING SAMPLES – V.N. Vimal Rao

Attached are three recent papers that are representative of my work over the past few years as a graduate student at the University of Minnesota.

Categorical Perception of p-values is the first paper in my primary line of research examining statistical cognition, specifically towards the development of a theory of a statistically situated numerical cognition. This paper won the 2021 Cognitive Science Society Disciplinary Diversity and Integration Award, and was subsequently published in Topics in Cognitive Science, Volume 14 Issue 2.

Graduate Students' Effect Size Category Boundaries builds off of the work conducted in *Categorical Perception of p-values* by investigating the way graduate students process and categorize effect sizes. This paper was presented at the 2022 International Conference on Teaching Statistics and will subsequently appear in the conference's published proceedings.

Interpretations and Uses of the Comprehensive Assessment of Outcomes in Statistics is an unpublished manuscript written to complete the preliminary written examination in the Quantitative Methods in Education program in the Department of Educational Psychology. This paper was written alone in my second year in the program. This paper is the basis of a manuscript in preparation, in which I collaborate and co-author the paper with the College of Education and Human Development's librarian to conduct a formal scoping review, and with the assistance of two other students in the charting and coding of manuscripts included in the review.

I have also attached my dissertation abstract to these three papers. Together, all of these papers represent the wide variety of methods and theories I use in the study of the psychology of statistics.

TOPICS
TOPICS IN COGNITIVE SCIENCE

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

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

Categorical Perception of *p*-Values

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

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

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

Abstract

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

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

1. Categorical perception of *p*-values

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

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

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

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

2. Background

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

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

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

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

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

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

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

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

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

Moreover, the MNL appears to be a logarithmically compressed distortion of the linear continuum of mathematics. The evidence for this comes in part from experiments where people are asked to identify the greater (or lesser) of a pair of numbers. People make faster judgments when the numbers are small versus large (e.g., 1 vs. 2 is faster than 8 vs. 9; Parkman, 1971); this is the *size effect*. They make faster judgments when the distance between numbers is large versus small (e.g., 2 vs. 8 is faster than 3 vs. 5; Moyer & Landauer, 1967); this is the *distance effect*. Finally, they make faster judgments for pairs of multidigit numbers when the digit in each place of the larger number is greater than its counterpart in the smaller number (e.g., 46 vs. 35 is faster than 45 vs. 36; Nuerk, Weger, & Willmes, 2001; Varma & Karl, 2013); this is the *compatibility effect*. There are also discontinuities of the MNL caused by the place-value symbol system for naming numbers. For example, determining the midpoint between two numbers is slower and less accurate when the tens digits differ (e.g., bisecting 27–35 is harder than bisecting 21–29; Nuerk, Moeller, Klein, Willmes, & Fischer, 2011); this is the *decade-crossing effect*.

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

3. Method

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

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

3.1. Participants

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

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

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

3.2. Materials

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

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

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

3.3. Procedure

Pairs of *p*-values were presented sequentially to participants, with the first *p*-value shown for 1000 ms and the second *p*-value shown until either a response was made or 5000 ms elapsed. Participants were asked to "identify whether the *p*-values are similar or different," and to indicate their choice as quickly as possible by pressing either "F" (for similar) or "J" (for different) on their keyboard. In approximately half of the trials, the first *p*-value presented was smaller than the second one, while in the other half it was larger. To induce a statistical mindset, *p*-values were presented in the form "$p = .0\text{AB}$." Stimuli were blocked into six sets of 18 pairs, with a 20 seconds break between each block. An additional six stimuli were presented in an initial training phase to familiarize participants with the task procedures.

3.4. Statistical methods

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

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

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

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

4. Results

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

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

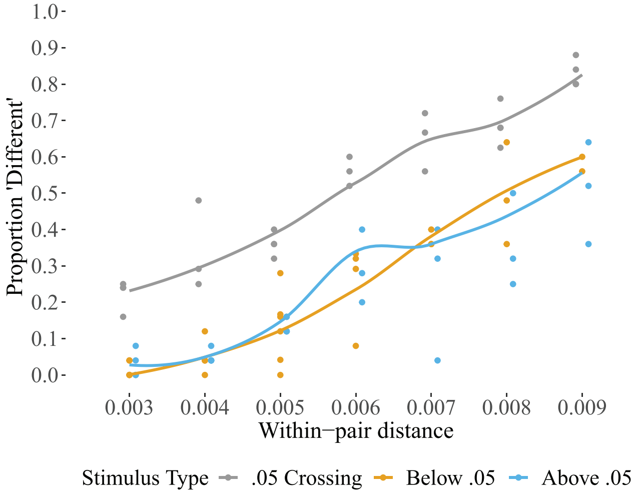

Fig. 1. Unadjusted proportion of participants selecting "different" per stimulus pair (dot) by within-pair distance (x-axis) and stimulus type (color), with weighted LOESS line.

Table 2
Fitted statistical model estimated fixed effects

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

[a] Estimates are per 0.001 increase.

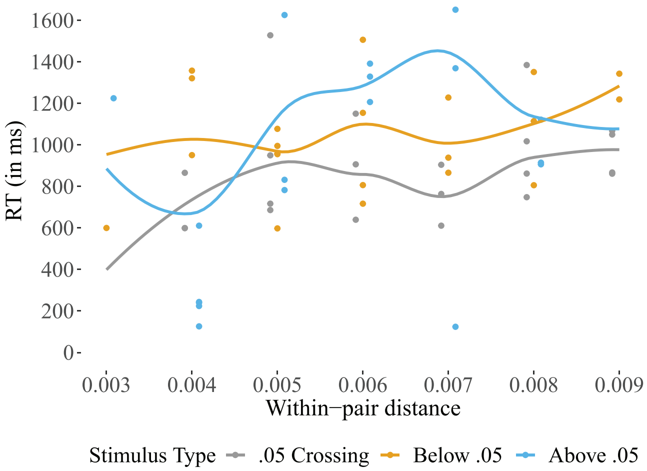

Fig. 2. Unadjusted mean RT for participants' "different" selections per stimulus pair (dot) by within-pair distance (x-axis) and stimulus type (color), with weighted LOESS line.

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

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

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

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

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

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

5. Discussion

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

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

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

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

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

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

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

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

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

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

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

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.

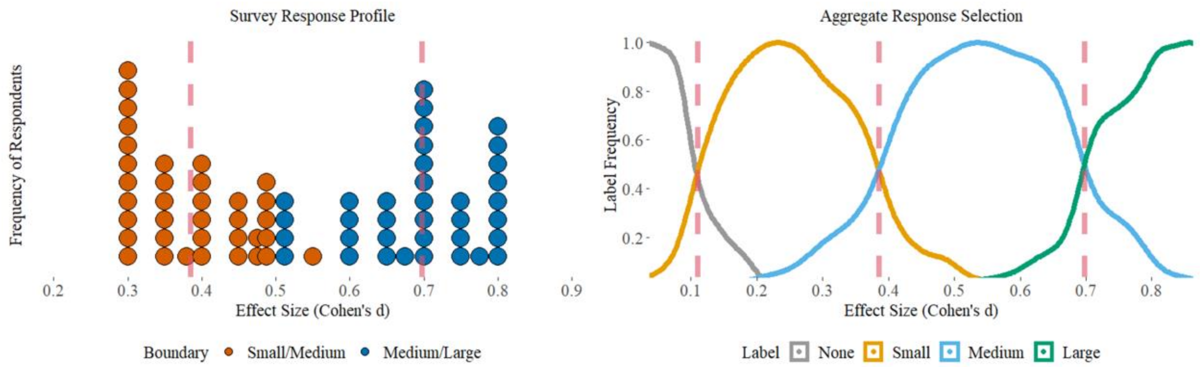

Figure 1: *Aggregate response patterns – Boundary location survey responses with reference lines at aggregate boundaries (left) and boundary identification task responses with reference lines at aggregate boundaries (right)*

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.

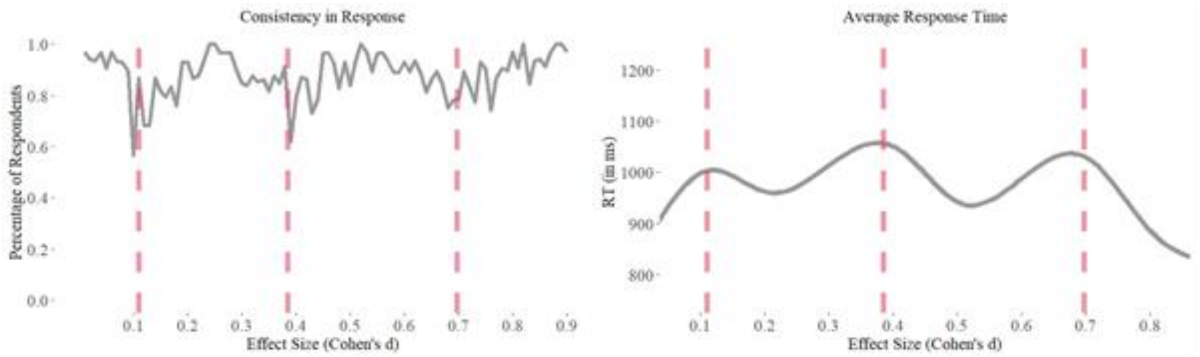

Figure 2: *Response patterns – Consistency in response selections and average response time by effect size value with reference lines at aggregate boundaries*

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.

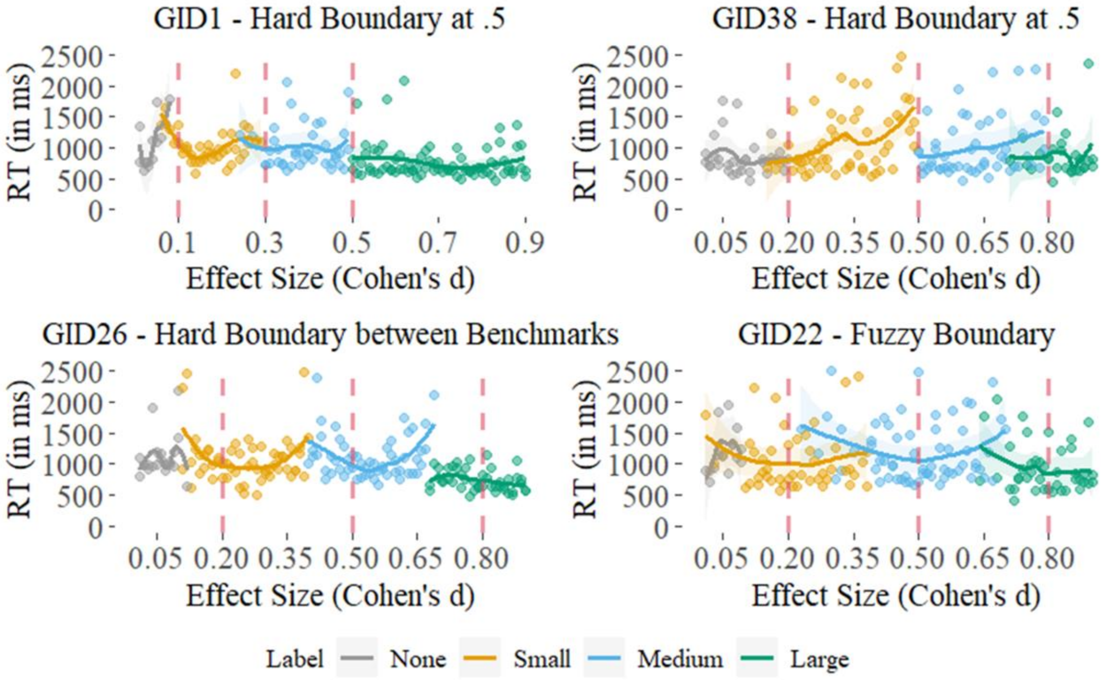

Figure 3: *Selected participant response patterns – Participants' responses (label) and response time (RT) with reference lines at common benchmark values as provided by Cohen (1988)*

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.

INTERPRETATIONS AND USES OF THE COMPREHENSIVE ASSESSMENT OF OUTCOMES IN STATISTICS

V.N. VIMAL RAO
University of Minnesota
rao00013@umn.edu

ABSTRACT

The Comprehensive Assessment of Outcomes in Statistics (CAOS) aims to measure students' conceptual understanding in statistics. A review of how CAOS scores have been interpreted for specific uses identified five themes: comparison of curricula, comparison of course formats, assessment and comparison of unique populations, identification of individual differences, and identification of relationships with other constructs. Few researchers explicitly included validity arguments supporting their interpretation of CAOS scores for their unique use, some of which may not be appropriate. CAOS users should ensure their proposed score interpretations for each intended use are supported by a preponderance of validity evidence and explicitly justified with a validity argument.

Keywords: *Statistics education; educational assessment; validity theory*

1. INTRODUCTION

The Comprehensive Assessment of Outcomes in Statistics (CAOS) is one of the most widely used assessments in statistics education research. CAOS was developed to measure student learning and conceptual understanding after completion of a tertiary-level first course in statistics (delMas et al., 2007). An evaluation of the test's content and a later evaluation of its internal structure (delMas, 2014) supported this intended score interpretation and test use.

CAOS was recently hailed as the 'gold standard' instrument to assess conceptual knowledge of statistics (Tintle & VanderStoep, 2018). It has been used in statistics education research to compare curricula and course formats (e.g., Ryan et al., 2016), to assess unique student populations (e.g., Fabrizio et al., 2011), and to identify associations with other constructs such as statistics attitudes and anxiety (e.g., Zonnefeld, 2015). It has also served as a blueprint for the development of new assessments (e.g., Chance et al., 2016).

While developing CAOS, delMas et al. (2007) envisioned several score interpretations for test uses and collected validity evidence to support CAOS's use for these purposes. Validity is a characteristic of particular interpretations for specified uses, each requiring their own validity arguments supported by validity evidence (APA, AERA, and NCME, 2014). Despite its wide application, few researchers utilizing CAOS have gathered validity evidence to support their unique uses, even when those uses differed from the original intended uses.

To heed and conform to the standards for educational testing and validity theory, the purpose of this study is to identify how CAOS scores have been interpreted and how CAOS has been used in statistics education research. Subsequently, the appropriateness of select unforeseen interpretations and uses will be briefly evaluated through the argument-based approach to validity (Kane, 2013; Sireci, 2013).

2. BACKGROUND

The Standards for Educational and Psychological Testing (APA, AERA, & NCME, 2014) provide guidance on the design, administration, and scoring of educational tests. In general, they ask researchers to pause and reflect on their methods. For example, CAOS users must be able to answer questions such as "How do I know CAOS actually is measuring what I want it to measure?" before analyzing CAOS results. These questions, and their answers, lie at the heart of *validity*.

2.1. VALIDITY

The *Standards* define validity as the "degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (APA, AERA, & NCME, 2014, p. 11). Perhaps more intuitive than this textbook definition, validity can simply be thought of as a test's appropriateness for a particular purpose (Kane, 2013; Sireci & Sukin, 2013). Tests are almost always administered to make certain decisions or claims (e.g., determining what students understand from a first course in statistics), and validity seeks to ensure that the interpretations of test scores are appropriate pieces of evidence to support such decisions or claims.

Four fundamental aspects of validity identified by Sireci (2007) are:

(1) Validity is a property of a specific purpose of a test, not the test itself. The *Standards* go so far as to state that "it is incorrect to use the unqualified phrase 'the validity of the test" (APA, AERA, & NCME, 2014, p. 11).

(2) Establishing validity requires multiple sources of evidence.

(3) A particular purpose must be defended by a "preponderance of evidence" (Sireci & Sukin, 2013, p. 64).

(4) Evaluating validity is a continual process. The defense of a proposed use or interpretation of a test is called *validation* and the main tool of validation is a *validity argument*.

A validity argument generally contains two components: how a test will be used or how its scores will be interpreted, and the evidence and logic that justifies each use and interpretation (APA, AERA, & NCME, 2014; Sireci & Sukin, 2013). Ideally, a validity argument establishes the plausibility and appropriateness of specific purposes and includes arguments both for and against the proposed use or interpretation (Cronbach, 1988; Messick, 1989). During test development, validity arguments can also help identify a need to refine or revise the test (APA, AERA, & NCME, 2014).

The *Standards* advocate the *argument-based approach* to validity (Kane, 2006). The three main steps to the argument-based approach and the development of a validation plan are, as summarized by Sireci (2013), are:

(1) The clear articulation of testing purposes. This includes intended test use, score interpretation, and inferences to be made based on test scores.

(2) Considerations of potential test misuse, including the *unintended social consequences* (Messick, 1989) of the interpretations of test scores.

(3) Evaluating test purposes and potential misuses with *validity evidence*.

Thorough and sound validity arguments incorporate many pieces of evidence as well as many sources of evidence. While there is no such thing as too much validity evidence, a practical threshold for determining sufficiency is when the intended interpretation or use is supported by a preponderance of evidence or when the positive consequences outweigh the negative consequences (Sireci, 2013, p. 104; Sireci & Sukin, 2013, p. 64).

Sireci and Sukin (2013) describe the amalgamation of validity evidence as akin to building a case in the courtroom. Lawyers do not limit themselves to one type of evidence to establish their case (Bex, 2011). Instead, they focus on building a comprehensive argument supported by multiple pieces of evidence to justify their claim. Furthermore, courtroom cases involve opposing attorneys each attempting to discredit the other's argument. Validation is not only essential in supporting an interpretation but also in finding out what may be wrong with it (Cronbach, 1980, p. 103). Researchers should adopt the role of both the prosecutor and defense attorney, attempting to discredit evidence in support of a validity argument and submitting evidence against the argument. Only after resilient evidence both for and against an argument have been considered can one determine that there is a preponderance of evidence in favor of a particular interpration of scores for a particular purpose. This burden of proof lies with the test user, and researchers must take the default position that an intended use is not valid until proven otherwise (APA, AERA, and NCME, 2014).

Types of sources of Validity Evidence The *Standards* specify five main types of sources of validity evidence based on test content, response processes, internal structure, relations with other variables, and testing consequences.

Evidence based on test content broadly refers to any analyses examining the relationship between the theoretical construct a test is attempting to measure and the specific content of the test (APA, AERA, and NCME, 2014). This type of source often includes expert judgment on the representativeness of items in relation to the construct of interest (Sireci & Faulkner-Bond, 2014). For example, evidence based on test content is particularly essential when validating certification exams.

Evidence based on response processes addresses the relationship between the construct and how test-takers interact with each item on the test (APA, AERA, and NCME, 2014). This type of source typically involves analyses of individuals' interaction with items, either through think-alouds, cognitive interviews, or analyses of their responses to items (Padilla & Benitez, 2014). This type of source is particularly important in ensuring that the question formats do not favor any one particular group of test takers.

Evidence based on internal structure are typically analyses examining the relationship between test items and the extent to which those relationships match the conceptual framework for the construct of interest. This type of source often includes examinations of dimensionality, measurement invariance, or reliability (Rios & Wells, 2014). Evidence based on internal structure is particularly essential to supporting interpretations of test scores or sub-scores. For example, if an assessment aims to measure both statistical thinking and computational thinking, there should be sufficient evidence to suggest that the items require the use of two distinct skill sets, and that the subscores contain distinct information (Haberman & Sinharay, 2010).

Evidence based on relations to other variables often includes comparisons of test scores to those from other tests that intend to measure similar constructs or to outcomes a test is purporting to predict. This also includes analyses of the discriminatory power of the test, in that it should be correlated to other related test scores, but should not be correlated to unrelated test scores (McCoach et al., 2013). This type of evidence is most often used to support claims that the assessment is consistent with the underlying construct.

Evidence based on testing consequences refer to the actions immediately following interpretations of test scores. Most tests are designed to drive change or make decisions in some form, and therefore, this type of evidence includes evaluations of the extent to which test results inform decision-making and the consequences of those decisions. This may include analysis of follow-up studies of individuals or cognitive interviews during the review of score reports (Lane, 2014). Not only are intended consequences important, but considerations of unintended consequences are essential forms of evidence that inform a validity argument. This type of evidence often is based on *value judgments* (Messick, 1989), and can support claims about the holistic value and results of a testing program. For example, a placement test whose scores are used to place students in different courses may inadvertently place an individual incorrectly. The consequences of this decision error should inform the validation of the test. More relevant to CAOS, Linn (2009) argues that assessments aimed to promote rigor in instruction should include evidence that changes to the level and depth of instruction occur as a result of test administration.

2.2. DESIGN AND DEVELOPMENT OF CAOS

The first step in the development of any test or assessment is the specification of the intended uses and interpretation of scores (Sireci, 2013). The creators of CAOS aimed to develop an assessment with reliable scores based on items that students completing any introductory statistics course would be expected to understand (delMas et al., 2007). Furthemore, they hoped to use the scores to identify areas where students improve, or fail to improve, in terms of their statistical understanding and reasoning. The next crucial step was the collection of evidence to support these uses and interpretations.

Before CAOS could measure what students know, delMas et al. (2007) had to decide what students should know at the end of a first course in statistics, i.e. the *content standards* (APA, AERA, and NCME, 2014, p.185). This process began by consulting the Assessment Resource Tools for Improving Statistical Thinking (ARTIST; Garfield et al., 2002) advisory group for advice and content validity ratings for an initial set of items selected from the ARTIST item database. A revised initial set of items was then administered as a pilot test to several students to ensure that items were functioning as intended, which resulted in either the omission or revision of several items.

After incorporating changes based on these validity ratings and the small field test, delMas et al. (2007) solicited validity ratings from statistics instructors before conducting a large field test. Data from this final test were provided to expert raters recruited from the advisory and editorial boards of the Consortium for the Advancement of Undergraduate Statistics Education. Unanimous agreement that CAOS measures appropriate outcomes and near unanimous agreement that the outcomes are common to most introductory level courses led delMas et al. (2007) to state that CAOS was “a valid measure of important learning outcomes in a first course in statistics” (p. 32). This evidence based on test content and response processes supported their claim that this *standards-based interpretation* of CAOS scores was appropriate (APA, AERA, and NCME, 2014).

Although CAOS covered many topic areas, it was primarily designed to focus on reasoning about variability. The extent to which all items on an assessment are measuring the same construct is called the test’s internal consistency (Davenport et al., 2015). While there are many ways to estimate internal consistency, the creators of CAOS focused on coefficient alpha (Cronbach, 1951). Raw scores from a sample of 1470 students yielded an estimated coefficient alpha of 0.82, prompting delMas et al. (2007) to judge CAOS as having “acceptable internal consistency” (p. 33).

A subsequent study by delMas (2014) re-examined CAOS’s internal structure, focusing on the test’s dimensionality. DelMas conducted a confirmatory factor analysis on 23,645 students’ scores on CAOS collected between 2005 and 2013. Results indicated that a unidimensional testlet model (Wainer et al., 2007) best fit the data. This evidence based on internal structure supported delMas's (2014) claim that "the CAOS test measures a single construct of statistical understanding of concepts covered in introductory statistics courses with sufficient internal measurement reliability for research purposes" (p. 6).

To identify gains in student understanding, delMas et al. (2007) administered CAOS as both a pretest and posttest to 763 students and measured changes in both total score as well as changes in individual items. Inferences about gains were made at the group level, analyzing the mean percentage point improvement in total scores, or the difference in mean proportion correct for individual items.

Standards 2.4 and 12.11 state that any analysis of differences between scores, such as pretest and posttest differences, and average scores at the group level, should be accompanied by estimates of reliability and precision, such as the standard error of the difference (APA, AERA, and NCME, 2014). Conforming to these standards, delMas et al. (2007) reported the standard error of the difference in group mean score between posttest and pretest of 0.433 percentage points (p. 34). DelMas (2014) analyzed the factor loadings of each item in a one-factor testlet model to assess item-total score correlation for all items. Although factor loadings did vary, all factor loadings were greater than 0.15. With a sample size of 763 students, this implies a standard error of measurement no higher than approximately 2.4 percentage points for the difference in average percentage correct between posttest and pretest. This evidence based on response processes and internal structure supported the claim that CAOS can be used to identify gains in students' understanding.

While delMas et al. (2007) only analyzed total scores and individual items to identify gains, based on the results of their analysis on individual items, they discuss results from each item "logically organized by topic areas" (p. 47). This organization can be traced to the origins of CAOS items from the ARTIST item bank, which has 11 topic scales (delMas et al., 2005). However, despite such a grouping, inferences were tied to specific items, and no subscore analysis was conducted. Furthermore, the identification of topic areas was related to a secondary purpose, which was to provide feedback to instructors and to promote changes in their instruction better aligned with the learning goals underlying CAOS (delMas et al., 2007, p. 50). Through interviews with statistics instructors, delMas et al. found that many instructors were surprised when they reviewed their students' scores on CAOS, indicating a potential gap between instructional content and emphases and those measured by CAOS. However, delMas et al. argue that this may be because many instructor-designed assessments have a focus on computation and formulas, while CAOS focuses on thinking and reasoning. Having seen their students' scores, many instructors reported that CAOS test results caused them to reflect on their teaching. This evidence based on test content and testing consequences supported delMas et al.'s proposed use of CAOS to drive instructional change.

2.3. SUBSEQUENT UTILIZATION

Although delMas et al. (2007) supported several intended uses of CAOS and interpretations of CAOS scores, validity is an ongoing process, and validity evidence is required for each utilization of an assessment (Kane, 2013; Sireci, 2007). However, it is not guaranteed that all CAOS users are familiar and have complied with the standards for educational testing.

While the *Standards* state that the responsibility for validating unintended test uses lies with the test user (APA, AERA, and NCME, 2014), Standard 12.16 states that:

> Those responsible for educational testing programs should provide appropriate training, documentation, and oversight so that the individuals who administer and score the test(s) are proficient in the appropriate test administration and scoring procedures and understand the importance of adhering to the directions provided by the test developer (p. 200).

To date, no comprehensive review of CAOS use has been completed, either by statistics education researchers or researchers in other fields. This study aims to rectify this fact by summarizing CAOS's use by statistics education researchers, and focuses on the following research questions:

(1) For what purposes has CAOS been considered for use?

(2) How have CAOS scores been interpreted?

Together, the answers to these questions can inform statistics education researchers' efforts to provide training and oversight to CAOS users and users of other assessments in statistics education research.

3. METHODS

To answer these research questions, this study utlized a scoping overview of literature to locate uses of CAOS combined with a thematic analysis to describe and summarize uses of CAOS.

3.1. SCOPING OVERVIEW

A scoping review is a rigorous method to collect and analyze data from a variety of sources (Arksey & O'Malley, 2005). Though similar to a systematic review, scoping reviews generally take a more exploratory nature and are useful when attempting to identify how research has been conducted on a certain topic (Munn et al., 2018). A scoring overview is similar to a scoping review, but is considered a less formal procedure in the initial investigation of a particular body of literature. General procedures for scoping reviews as outlined in the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) extension for Scoping Reviews Statement (Tricco et al., 2018) were used as a basis to identify uses of CAOS in statistics education research. Originally developed for the medical field, the PRISMA guidelines are currently used by many fields and have been previously used in statistics education research (e.g., Nolan et al., 2012).

This scoping overview only included only publications citing delMas et al. (2007). Two citation-based search tools were used to identify works citing delMas et al.: Web of Science and SCOPUS. Web of Science, maintained by Clarivate Analytics, is a citation database that includes research published in many high-impact journals and conference proceedings. Scopus, maintained by Elsevier, is a citation database that includes research published in books and journals.

As seen in Figure 1, a cited reference search in Web of Science yielded 71 results, including 56 articles and 11 conference papers. In SCOPUS, the search terms included "delMas" as author, "2007" as year, and any of the words "assessing conceptual understanding statistics" in any order as part of the title. This search returned 117 results, including 100 journal papers and eight conference papers.

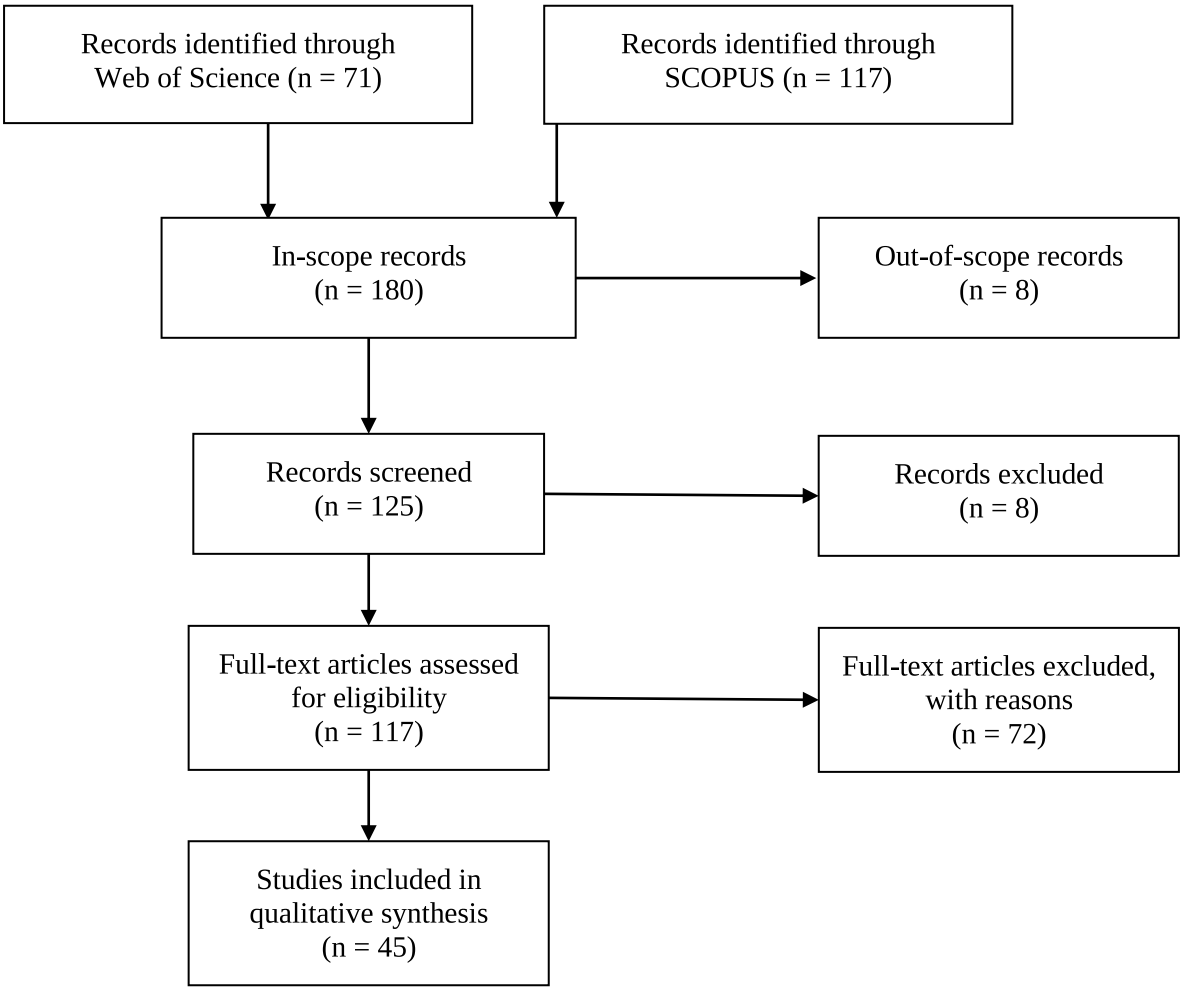
Records identified through Web of Science (n = 71)
Records identified through SCOPUS (n = 117)
In-scope records (n = 180)
Out-of-scope records (n = 8)
Records screened (n = 125)
Records excluded (n = 8)
Full-text articles assessed for eligibility (n = 117)
Full-text articles excluded, with reasons (n = 72)
Studies included in qualitative synthesis (n = 45)

Figure 1. Scoping review flow diagram

Across both sources there were 125 distinct results that cited delMas et al. (2007). Of these, eight were excluded due to access limitations. Of the remaining 117 results, 102 were journal papers, nine were conference papers, and six were book chapters. Of the 102 journal papers, 46 were included in both Web of Science and SCOPUS, nine were in Web of Science only, and 47 were in SCOPUS only. Many of the latter 47 search results were from the Statistics Education Research Journal and the Journal of Statistics Education, two of the pre-eminent journals in statistics education research. Of the nine conference papers, four were included in both databases, with an additional two from SCOPUS only and three from Web of Science only. Of the six citing book chapters, five were included in SCOPUS, while only two were included in Web of Science.

Every search result was initially reviewed by reading the abstract and finding and analyzing all in-text citations of delMas et al. (2007). While delMas et al. introduce CAOS in their paper, they also analyze results from their sample responses. Therefore, many citing works simply cited delMas et al. as part of a literature review without consideration of using CAOS. These citing works were subsequently excluded from further analysis. Of the remaining 45 citing works, 18 were included in both the Web of

Science results as well as the SCOPUS results, 22 were in SCOPUS only, and five were in Web of Science only. The latter five included two journal papers, two conference papers, and one book chapter.

For citing works that considered using CAOS, initial codes were assigned and stored to indicate study purpose, the CAOS use considered by the authors, and their argument for or against its use. If CAOS was used, claims and interpretations based on CAOS use, in the form of text excerpts, were extracted and stored in an electronic database. These reasons, justifications, interpretations, and claims were then examined based on a thematic analysis to summarize general patterns for both research that did use CAOS and research that did not use CAOS.

3.2. THEMATIC ANALYSIS

Thematic analysis is a rigorous method used to identify patterns by synthesizing a database including multiple sources of text into a few broad categories (Braun & Clarke, 2006). Although thematic analyses more typically involve the review of interview transcripts and observation notes, it has also been used in conjunction with systematic reviews to analyze the qualitative component of data generated in such processes (Thomas & Harden, 2008).

While coding processes and frameworks govern data collection and description in systematic reviews, they are operationalized slightly differently depending on the research framework (Auerbach & Silverstein, 2003). Thematic analysis utilizes an applied coding framework specifically designed to identify themes that has several variants. Each variant of thematic analysis specifies a slightly different relationship between codes and content. This study utilizes an inductive variant of the reflexive thematic analysis (Braun & Clarke, 2019).

In the inductive variant, codes and themes are fully generated by the content and data itself without any influence of outside theory. Such thematic analyses are conducted in six steps: familiarization, coding, generating themes, reviewing themes, defining themes, and naming themes. Although the steps are sequential, they do not constitute a rigid order that must be followed in sequence. Rather, thematic analysis is a recursive process with many steps either blending together as needed, akin to cyclic coding (Saldaña, 2015).

After all data were extracted, initial codes were assigned as succinct summarizations for each piece of information extracted, i.e., study purpose, CAOS use, use reason, score interpretation, validity argument, and validity evidence. There were four clear categories of CAOS use that quickly emerged: CAOS used in full without changes; a large subset of CAOS items were used in full without changes; a small subset of CAOS items were modified and used in combination with other items; or CAOS was not used at all. For the interpretation of scores, initial codes were assigned separately to interpretations of total scores, subscores, or performance on individual items, which resulted in a diverse set of codes. After codes were assigned, they were examined and collated into broad groups that represented potential themes. Each theme was then reviewed against the original data to ensure appropriateness of fit and consistency. This led to some of the initial groupings being split and others combined to achieve a consistent level of abstraction applied to the full data set and for each research question.

4. RESULTS

In general, the most common way CAOS was used by researchers was a guide or basis for the creation of new items or assessments. The most common interpretation of

CAOS scores when CAOS was used in its entirety was to interpret differences in scores between groups in quasi-experimental studies. The four main themes of considered uses of CAOS identified in this review are: content coverage, measurement reliability and level of detail, test developments, and one of delMas et al.'s (2007) intended uses. Across all of the 19 citing works using CAOS without modification, five major themes emerged describing how CAOS scores, either total score, subscores, or item scores, have been interpreted: a comparison of students' understanding across curricula, a comparison of students' understanding across course formats, the assessment and comparison of student understanding in unique populations, the identification of individual differences in understanding, and relationships with other constructs related to statistics education (see Table 1).

4.1. FOR WHAT PURPOSES HAS CAOS BEEN CONSIDERED FOR USE?

Of the 45 citing works that considered using CAOS, 12 ultimately chose not to use CAOS, choosing to utilize other assessments to measure understanding of statistics. This included the use of the Levels of Conceptual Understanding in Statistics (Lovett & Lee, 2018), the Statistical Reasoning Assessment (Gundlach et al., 2015; Martin et al., 2017), the Statistics Concept Inventory (Lauriski-Karriker et al., 2013; Richardson, 2011), and ARTIST topic scales (Castro Sotos et al., 2009; Monárrez et al., 2018).

A few researchers cited misalignment between their curricula and CAOS as the reason why they chose not to use CAOS (Callingham & Watson, 2017; Crooks et al., 2018; Spence et al., 2016). Others commented on the response process of CAOS being insufficient to capture information related to the researchers' goals, instead preferring constructed response items to CAOS's multiple choice items (De Vetten et al., 2019; Zimmerman et al., 2018). Some researchers commented on the estimated reliability of CAOS and judged the test to have insufficient reliability for their purposes or compared to another assessment (Fawcett, 2017; Olani et al., 2010).

A further 14 of the 45 citings works that considered using CAOS ultimately choose to only utilize a small subset of items in the development of an ad hoc assessment, often along with modifications. This was often accompanied with a validity argument citing evidence based on test content to explain why CAOS would not be appropriate for their uses, although none of the citing works explicitly labeled their arguments as such. The most commonly cited example of evidence based on test content was a determination that the statistics topics covered in the course the researcher wished to measure were not aligned with those covered by CAOS (e.g., Beckman et al., 2017; Chance et al., 2016). Similarly, researchers adjusted CAOS items due to differences in the context of their courses, such as re-writing items with contexts familiar to biology students (Corredor, 2012; Matthews et al., 2016; Stanhope et al., 2017). A smaller subset of researchers modified CAOS items to better measure different constructs related to statistical reasoning, such as reasoning associated with statistical modelling or statistical literacy (e.g., Vidic et al., 2014; Ziegler & Garfield, 2018).

All but two of the 19 citing works using CAOS without modification expressed an intended utilization roughly aligned with at least one of the four intended uses envisioned by delMas et al. (2007). Eight citing works used CAOS to measure what groups of students know about statistics (e.g., Duarte & Cazares, 2014; Hannigan et al., 2013). Ten citing works used CAOS to measure student growth (e.g., Groth & Bergner, 2013; Hahs-Vaughn et al., 2017). Three citing works used CAOS to identify differences in understanding or gains in understanding by topic area (Hildreth et al., 2018; Tintle et al., 2018; Wang et al., 2019).

Perhaps one of CAOS's most influential uses has been as a basis for the development of new assessments. Researchers have consistently looked to CAOS as a starting point in their own efforts. This has led to the development of assessments such as the Assessment of Inferential Reasoning in Statistics (AIRS; Park, 2012), the Goals and Outcomes Associated with Learning Statistics (GOALS; Garfield et al., 2012; Sabbag & Zieffler, 2015), the Statistical Reasoning in Biology Concept Inventory (SRBCI; Deane et al., 2016), the Quantitative Skills Assessment of Science Students (QSASS; Matthews et al., 2017), the Biology Science Quantitative Reasoning Exam (BioSQuaRE; Stanhope et al., 2017), and the Basic Literacy in Statistics assessment (BLIS; Ziegler & Garfield, 2018).

4.2. HOW HAVE CAOS SCORES BEEN INTERPRETED?

While delMas et al. (2007) describe CAOS as a measure of students' conceptual understanding, many citing works describe the constructs that they wish to measure using CAOS in different ways. These include probabilistic reasoning (Cao & Banaji, 2020), statistical literacy (Bowen et al., 2014; Hahs-Vaughn et al., 2017), and statistical reasoning and thinking (Conway IV et al., 2019; Tintle et al., 2012). However, previous research has indicated that there is little consensus in the nuances between the related constructs of understanding, reasoning, thinking, and literacy (delMas, 2004). Furthermore, attempts to measure distinct aspects of these constructs have generally failed to find sufficient evidence of multidimensionality (e.g., Sabbag et al., 2018).

Some researchers have used CAOS to compare student understanding across cohorts. This has included studies comparing curricula, e.g., simulation based inference curricula (Hildreth et al., 2018; Tintle et al., 2018), and studies comparing novel course formats, e.g., online and hybrid formats (e.g., Conway IV et al., 2019; Posner, 2011). These comparisons have included interpretations of differences in total scores, subscores by topic, and scores by item between groups as well as group differences in student growth between pretest and posttest administration. Interpretations of total scores have included statements interpreting differences in means scores in terms of higher performance in one of the groups (e.g., Bowen et al., 2014; Tintle et al., 2018). Interpretations of subscores included statements such as specifying the number of subscales in which one group outperformed the other (Tintle et al., 2018). Differences in item-level performance were interpreted in terms of adjusted odds ratios for answering correctly between the two curricula being compared (Hildreth et al., 2018). These interpretations led to claims that, for example, the SBI curriculum is beneficial for students (Hildreth et al., 2018), or that online courses do not negatively impact student learning (Bowen et al., 2014).

One of the most common uses of CAOS is to assess unique student populations. This includes students from countries other than the United States (Duarte & Cazares, 2014; Saputra et al., 2018), graduate students (Hahs-Vaughn, 2017; Wang et al., 2019), pre-service teachers (Groth & Bergner, 2013; Hannigan et al., 2013), and students at different levels of undergraduate training (Horton, 2013; Chance & Peck, 2015). Calculations of the average total scores often were accompanied with comparisons to the results reported in delMas et al. (2007) (e.g., Duarte & Cazares, 2014; Hahs-Vaughn et al., 2017) or between groups of students (Lübke et al., 2019; Hannigan et al., 2013). The most common interpretation of total scores was a statement of the average level of conceptual understanding of students in the population of interest (e.g., Hannigan et al., 2013; Horton, 2013) or their growth (e.g., Saputra et al., 2018). One study utilized CAOS sub-scores to assess a unique student population. Wang et al. (2019) used a subset of items related to confidence intervals to measure graduate students' understanding, calculating students' total score on this subset of items, ipso facto calculating a CAOS subscore.

Table 1. Interpretations and Uses of CAOS scores in uses of CAOS without modification

Abbreviated Citation	Total score use	Sub score use	Item score use
Posner (2011)	Comparing course format, describing student populations, relationships with other constructs		
Tintle et al. (2011)	Comparing curricula, pre- and post-		Comparing curricula, pre- and post-
Tintle et al. (2012)	Comparing curricula, pre- and post-	Comparing curricula, pre- and post-	Comparing curricula, pre- and post-
Groth & Bergner (2013)	Describing student populations, pre- and post-		
Hannigan et al. (2013)	Describing student populations		Describing student populations
Horton (2013)	Pre- and post-		
Leavy et al. (2013)	Relationships with other constructs		
Bowen et al. (2014)	Comparing course format, pre- and post-, relationships with other constructs		
Duarte & Cazares (2014)			Describing student populations
Fitzmaurice et al. (2014)	Relationships with other constructs		
Chance & Peck (2015)	Pre- and post-		
Hahs-Vaughn et al. (2017)	Comparing course format, describing student populations, pre- and post-, relationships with other constructs		
Hildreth et al. (2018)			Comparing curricula
Saputra et al. (2018)	Pre- and post-	Pre- and post-	
Tintle et al. (2018)	Comparing curricula, describing student populations, pre- and post-, relationships with other constructs	Comparing curricula, describing student populations, pre- and post-	
Cao & Banaji (2020)	Describing student populations		
Conway IV et al. (2019)	Comparing course format, relationships with other constructs		
Lübke et al. (2019)		Describing student populations, pre- and post-	
Wang et al. (2019)		Pre- and post-	Pre- and post-

Two studies utilized CAOS to separate students based on their level of conceptual understanding. Cao & Banaji (2020) utilized CAOS to separate participants into groups of higher and lower probabilistic reasoning ability. These groups were then used to analyze students' tendencies in a custom task to assess estimation bias. Similarly, Tintle et al. (2018) utilized CAOS to separate students into groups of low, medium, and high levels of conceptual understanding of statistics. Tintle et al. used these groups as a covariate for assessing student growth between two different curricula. CAOS total scores and subscores were then interpreted separately for each group, leading to statements about the improvement of each group within each subtopic of CAOS.

A few researchers have also used CAOS to assess the relationship between conceptual understanding of statistics and other related constructs. This has included the use of CAOS in conjunction with students' attitudes towards statistics (e.g., Fitzmaurice et al., 2014; Posner, 2011) and teacher and classroom characteristics (Bowen et al., 2014, Conway IV et al., 2019). These uses do not typically carry separate interpretations of CAOS total scores. Rather, CAOS scores are analyzed as part of a model, and its relationship with other constructs are interpreted in terms of correlation coefficients and model fit. For example, Conway IV et al. (2019) interpret the η^2 statistic when assessing the relationship between variation in teacher characteristics and variation in students' conceptual understanding. Similarly, Leavy et al. (2013) use the *r* statistic when assessing the relationship between students' conceptual understanding and their attitudes towards statistics.

5. DISCUSSION

This study aimed to provide an overview of CAOS uses and score interpretations by statistics education researchers. The most common CAOS use by researchers was a guide or basis for the creation of new items or assessments. The most common interpretation of CAOS scores when CAOS was used without modification was to examine differences in scores between groups in quasi-experimental studies (e.g., students in different curricula or course formats).

Although each use of an assessment and interpretation of scores should be accompanied with its own validity evidence, only a few researchers using CAOS explicitly provided or implicitly alluded to a validity argument supporting their use or collected new validity evidence for their proposed use. Hannigan et al. (2013) justify their use by noting that the content of the statistics course for which they desired to use CAOS aligns with the content of CAOS. Tintle et al. (2012) assessed the internal consistency of CAOS scores from their participants. They estimated the coefficient alpha (Cronbach, 1951) and judged that although their estimate was lower than that of delMas et al. (2007), it still met an acceptable level of reliability. Bowen et al. (2014) included participant characteristics such as gender and race as covariates in a model assessing differences in conceptual understanding based on course format, a tacit nod to potential measurement invariance.

The absence of validity arguments with a preponderance of evidence supporting each unique use and score interpretation of CAOS leaves open a question as to their appropriateness. In particular, two interpretations of scores and uses stand out as potentially invalid: the calculation of subscores and the assessment and comparison of unique populations.

While CAOS was designed to cover multiple content topics, and although delMas et al. (2007) grouped their interpretations of CAOS results by items into topics, the calculation of CAOS subscores requires additional validation in order to meet thresholds for *distinctness* and *reliability* (APA, AERA, and NCME, 2014, p. 27). In general, subscores based on content subdomains are not often recommended for research purposes, typically due to low levels of reliability when the number of items is small (Biancarosa et al., 2019; Sinharay et al., 2007). A preliminary analysis by Rao and Chavez (2020) utilizing CAOS data collected between 2007 and 2019 concluded that CAOS subscores, as defined by content topic, were neither reliable nor distinct. Therefore, based on this evidence of the internal structure of CAOS, it does not appear that the use of CAOS subscores is appropriate.

Despite the fact that CAOS was tested with a sample of students across the United States, this sample was not randomly selected, nor considered by delMas et al. (2007) to be representative of all students in the nation. Therefore, the total scores as reported by delMas et al. cannot be construed as a norm. Differences in overall performance between samples may simply be due to each sample's differing constituency. This is particularly crucial to consider as no study to date has comprehensively assessed CAOS for measurement invariance. In an evaluation of ARTIST items, Monárrez et al. (2018) found that english language learners experience difficulties answering particularly context-laden questions, compared to native english speakers (p. 171). This is especially problematic when CAOS is used for non-native english speakers. Therefore, without sufficient evidence of the responses processes or internal structure of CAOS to ensure that CAOS functions identically for different groups of individuals, it does not appear that the use of CAOS to compare unique student populations is appropriate.

5.1. IMPLICATIONS

While this review focused on uses of CAOS for research purposes and limited its inquiry to publications citing delMas et al., 2007, all potential users of CAOS should carefully weigh the appropriateness of their intended score interpretations and proposed uses. For example, classroom instructors should consider the reliability and distinctness of subscores before deciding to alter the structure of their curriculum to place greater focus on a topic based on average CAOS subscores.

Researchers planning on using CAOS, or another assessment, should explicitly justify each proposed score interpretation for a particular use with a validity argument based on validity evidence. The process of creating such validity arguments, in conformance with the standards for educational and psychological testing, will help to ensure rigor in conclusions drawn from statistics education research.

Assessment specialists should support the research community by conducting research to garner potential validity evidence to support intended score interpretations and uses of CAOS, or other assessments. For example, investigations of the internal structure of assessments is important validity evidence to allow researchers to compare different student populations, such as students across countries, institutions, or disciplines.

Finally, as a scoping overview, this study is not a comprehensive review of all uses of CAOS in research. However, it provides a platform for future reviews to systematically examine the manner in which CAOS has been used by researchers, and provides a framework for such investigations.

5.2. SUMMARY

In 2007, delMas et al. introduced the statistics education community to the Comprehensive Assessment of Outcomes in Statistics (CAOS). They intended for CAOS to be used to measure what students in a first course in statistics know compared to the expectations for their conceptual understanding. It was also intended to measure student growth from the start to the end of the course. They subsequently expected these scores to be used by instructors to reflect on their teaching. To facilitate these score interpretations, delMas et al. designed CAOS to have reliable scores.

Since its introduction, CAOS has become one of the most widely used assessments in statistics education research at the post-secondary level (Tintle & VanderStoep, 2018). Many researchers have used CAOS in full or in part, in addition to using CAOS as a base for the development of new assessments. However, many researchers have interpreted CAOS scores for unique uses different from those originally envisioned by delMas et al. (2007). This includes the comparison of curricula and course format and the assessment of unique student populations (e.g., Hildreth et al., 2014; Tintle et al., 2018). CAOS scores have also been calculated and used in ways unintended by delMas et al. Researchers have calculated subscores by topic (e.g., Lübke et al., 2019) and used scores to identify individual differences in conceptual understanding of statistics (e.g., Cao & Banaji, 2020).

Few of these novel uses or score interpretations were explicitly supported by validity arguments, thus leaving open the question of their appropriateness. Preliminary investigations suggest that CAOS subscores by topic are neither distinct nor reliable, casting doubt on the validity of the interpretation of subscores. Similarly, analyses suggest there may be measurement invariance based on individual characteristics such as gender and race/ethnicity as well as institution type. This suggests that interpretations of scores to identify individual differences, compare student populations, or to compare course formats may not be valid.

Each proposed use of CAOS and proposed interpretation of CAOS scores for research purposes should be accompanied with a validity argument supported by a preponderance of validity evidence. Ingraining the *Standards for Educational and Psychological Testing* into the assessment practices of the statistics education research community will help to ensure that claims we make and actions we take based on CAOS are ones that are appropriate and consistent with standards of educational measurement.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International journal of social research methodology, 8*(1), 19-32.

Auerbach, C., & Silverstein, L. B. (2003). *Qualitative data: An introduction to coding and analysis* (Vol. 21). NYU press.

Beckman, M. D., Delmas, R. C., & Garfield, J. B. (2017). Cognitive Transfer Outcomes for a Simulation-Based Introductory Statistics Curriculum. *Statistics Education Research Journal, 16*(2), 419-440.

Bex, F. J. (2011). *Arguments, stories and criminal evidence: A formal hybrid theory* (Vol. 92). Springer Science & Business Media.

Biancarosa, G., Kennedy, P. C., Carlson, S. E., Yoon, H., Seipel, B., Liu, B., & Davison, M. L. (2019). Constructing subscores that add validity: A case study of identifying students at risk. *Educational and psychological measurement*, *79*(1), 65-84.

Bowen, W. G., Chingos, M. M., Lack, K. A., & Nygren, T. I. (2014). Interactive learning online at public universities: Evidence from a six-campus randomized trial. *Journal of Policy Analysis and Management, 33*(1), 94-111.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology, 3*(2), 77-101.

Braun, V., & Clarke., V. (2019, April) *Thematic analysis: a reflexive approach.* The University of Auckland. https://www.psych.auckland.ac.nz/en/about/thematic-analysis.html

Callingham, R., & Watson, J. M. (2017). The Development of Statistical Literacy at School. *Statistics Education Research Journal, 16*(1), 181-201.

Cao, J., & Banaji, M. R. (2020). Inferring an unobservable population size from observable samples. *Memory & Cognition, 48*(3), 348-360.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests?. *Journal of Statistics Education, 17*(2), 1-21.

Chance, B., & Peck, R. (2015). From curriculum guidelines to learning outcomes: Assessment at the program level. *The American Statistician, 69*(4), 409-416.

Chance, B., Wong, J., & Tintle, N. (2016). Student performance in curricula centered on simulation-based inference: A preliminary report. *Journal of Statistics Education, 24*(3), 114-126.

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum?. *Technology innovations in statistics education, 1*(1), 1-15.

Conway IV, B., Gary Martin, W., Strutchens, M., Kraska, M., & Huang, H. (2019). The Statistical Reasoning Learning Environment: A Comparison of Students' Statistical Reasoning Ability. *Journal of Statistics Education*, *27*(3), 171-187.

Corredor, J. A. (2012). Effects of the Amount of Activity on the Learning of Data Analysis and Sampling Distribution in the Context of Statistics Teaching: An Imperfect Comparison. *Revista Colombiana de Psicología, 21*(2), 285-302.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cronbach, L. J. (1980). Validity on parole: How can we go straight. *New directions for testing and measurement, 5*(1), 99-108.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Crooks, N. M., Bartel, A. N., & Alibali, M. W. (2019). Conceptual Knowledge of Confidence Intervals in Psychology Undergraduate and Graduate Students. *Statistics Education Research Journal, 18*(1), 46-62.

Davenport, E. C., Davison, M. L., Liou, P. Y., & Love, Q. U. (2015). Reliability, dimensionality, and internal consistency as defined by Cronbach: Distinct albeit related concepts. *Educational Measurement: Issues and Practice, 34*(4), 4-9.

de Vetten, A., Schoonenboom, J., Keijzer, R., & van Oers, B. (2019). Pre-service primary school teachers' knowledge of informal statistical inference. *Journal of Mathematics Teacher Education, 22*(6), 639-661.

Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2016). Development of the statistical reasoning in biology concept inventory (SRBCI). *CBE—Life Sciences Education, 15*(1), ar5.

delMas, R. C. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 79-95). Dordrecht, The Netherlands: Kluwer Academic Publishers.

delMas, R. C. (2014). Trends in students' conceptual understanding of statistics. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

delMas, R. C., Garfield, J. B., & Ooms, A. (2005, July). Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. In K. Makar (Ed.), *Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy* (on cd). Auckland, New Zealand.

delMas, R. C., Garfield, J. B., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28-58.

Duarte, J. A. J., & Cazares, S. I. (2014). Comprensión y razonamiento de profesores de Matemáticas de bachillerato sobre conceptos estadísticos básicos [Mathematics teachers' comprehension and reasoning of basic statistics concepts]. *Perfiles educativos, 36*(146), 14-29.

Fabrizio, M., López, M. V., & Plencovich, M. C. (2011). Statistics in teacher training colleges in Buenos Aires, Argentina: Assessment and challenges. In *Proceedings of the 56th Session of the International Statistics Institute.* Lisbon: Portugal.

Fawcett, L. (2017). The CASE Project: Evaluation of case-based approaches to learning and teaching in statistics service courses. *Journal of Statistics Education, 25*(2), 79-89.

Fitzmaurice, O., Leavy, A., & Hannigan, A. (2014). Why Is Statistics Perceived as Difficult and Can Practice during Training Change Perceptions? Insights from a Prospective Mathematics Teacher. *Teaching Mathematics and Its Applications, 33*(4), 230-248.

Garfield, J. B., delMas, R. C., & Chance, B. (2002). *The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project.* NSF CCLI grant ASA- 0206571.

Garfield, J. B., delMas, R. C., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM, 44*(7), 883-898.

Groth, R. E., & Bergner, J. A. (2013). Mapping the structure of knowledge for teaching nominal categorical data analysis. *Educational Studies in Mathematics, 83*(2), 247-265.

Gundlach, E., Richards, K. A. R., Nelson, D., & Levesque-Bristol, C. (2015). A comparison of student attitudes, statistical reasoning, performance, and perceptions for web-augmented traditional, fully online, and flipped sections of a statistical literacy class. *Journal of Statistics Education, 23*(1), 1-23.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75*(2), 209-227.

Hahs-Vaughn, D. L., Acquaye, H., Griffith, M. D., Jo, H., Matthews, K., & Acharya, P. (2017). Statistical literacy as a function of online versus hybrid course delivery format for an introductory graduate statistics course. *Journal of Statistics Education, 25*(3), 112-121.

Hannigan, A., Gill, O., & Leavy, A. M. (2013). An investigation of prospective secondary mathematics teachers' conceptual knowledge of and attitudes towards statistics. *Journal of Mathematics Teacher Education, 16*(6), 427-449.

Hildreth, L. A., Robison-Cox, J., & Schmidt, J. (2018). Comparing Student Success and Understanding in Introductory Statistics under Consensus and Simulation-Based Curricula. *Statistics Education Research Journal, 17*(1), 103-120.

Horton, N. J. (2013). I hear, I forget. I do, I understand: a modified Moore-method mathematical statistics course. *The American Statistician, 67*(4), 219-228.

Jacob, B., Lee, H., Tran, D., & Doerr, H. (2015, February). Improving teachers' reasoning about sampling variability: A cross institutional effort. *CERME 9 - Ninth Congress of the European Society for Research in Mathematics Education*. Prague, Czech Republic

Kane, M. T. (2006). Validation. In B.L. Robert (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Wesport, CT: Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*(1), 127-135.

Lauriski-Karriker, T., Nicoletti, E., & Moskal, B. (2013). Tablet computers and inksurvey software in a college engineering statistics course: How are students' learning and attitudes impacted? *The ASEE Computers in Education (CoED) Journal, 4*(1), 43-50.

Leavy, A. M., Hannigan, A., & Fitzmaurice, O. (2013). If you're doubting yourself then, what's the fun in that? An exploration of why prospective secondary mathematics teachers perceive statistics as difficult. *Journal of Statistics Education, 21*(3), 1-25.

Linn, R.L. (2009). The concept of validity in the context of NCLB. In R. Lissitz (Ed.), *The concept of validity* (pp. 195-212). Charlotte, NC: Information Age Publishers

Lovett, J. N., & Lee, H. S. (2018). Preservice secondary mathematics teachers' statistical knowledge: A snapshot of strengths and weaknesses. *Journal of Statistics Education, 26*(3), 214-222.

Lübke, K., Gehrke, M., & Markgraf, N. (2019). Statistical Computing and Data Science in Introductory Statistics. In *Applications in Statistical Computing* (pp. 139-150). Springer, Cham.

Martin, N., Hughes, J., & Fugelsang, J. (2017). The Roles of Experience, Gender, and Individual Differences in Statistical Reasoning. *Statistics Education Research Journal, 16*(2), 454-475.

Matthews, K. E., Adams, P., & Goos, M. (2016). Quantitative skills as a graduate learning outcome of university science degree programmes: student performance explored through the planned–enacted–experienced curriculum model. *International Journal of Science Education, 38*(11), 1785-1799.

Matthews, K. E., Adams, P., & Goos, M. (2017). Quantitative skills as a graduate learning outcome: exploring students' evaluative expertise. *Assessment & Evaluation in Higher Education, 42*(4), 564-579.

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). Evidence based on relations to other variables: Bolstering the empirical validity arguments for constructs. In *Instrument development in the affective domain* (pp. 209-248). Springer, New York, NY.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Old Tappan, NJ: Macmillan.

Monárrez, A., Galvan, L., Wagler, A. E., & Lesser, L. M. (2018). Range of Meanings: A Sequential Mixed Methods Study of How English Language Learners Encounter Assessment Items on Descriptive Statistics. *Journal of Statistics Education, 26*(3), 162-173.

Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology, 18*(1), 143.

Nolan, M. M., Beran, T., Hecker, K. G. (2012). Surveys assessing students' attitudes toward statistics: A systematic review of validity and reliability. *Statistics Education Research Journal, 11*(2), 103-123.

Olani, A., Harskamp, E., Hoekstra, R., & van der Werf, G. (2010). The roles of self-efficacy and perceived teacher support in the acquisition of statistical reasoning abilities: a path analysis. *Educational Research and Evaluation, 16*(6), 517-528.

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136-144.

Park, J. (2012). *Developing and Validating an Instrument to Measure College Students' Inferential Reasoning in Statistics: An Argument-Based Approach to Validation.* (Doctoral dissertation, University of Minnesota).

Posner, M. A. (2011). The impact of a proficiency-based assessment and reassessment of learning outcomes system on student achievement and attitudes. *Statistics Education Research Journal, 10*(1), 3-15.

Rao, V.N.V., & Chavez, C. (2020, February). *On the Utilization of the Comprehensive Assessment of Outcomes in Statistics.* Poster presented at the 2020 Department of Educational Psychology Graduate Student Research Day, Minneapolis, Minnesota.

Richardson, A. M. (2012). Clickers in a First Statistics Course. In *Sustainable Language Support Practices in Science Education: Technologies and Solutions* (pp. 195-225). IGI Global.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108-116.

Ryan, S., Kaufman, J., Greenhouse, J., She, R., & Shi, J. (2016). The effectiveness of blended online learning courses at the community college level. *Community College Journal of Research and Practice, 40*(4), 285-298.

Sabbag, A., Garfield, J. B., & Zieffler, A. (2018). Assessing statistical literacy and statistical reasoning. *Statistics Education Research Journal, 17*(2), 141-160.

Sabbag, A. G., & Zieffler, A. (2015). Assessing learning outcomes: An analysis of the GOALS-2 instrument. *Statistics Education Research Journal, 14*(2), 93-116.

Saldaña, J. (2015). *The coding manual for qualitative researchers*. Sage.

Saputra, K. V. I., Cahyadi, L., & Sembiring, U. A. (2018, January). Assessment of statistical education in Indonesia: Preliminary results and initiation to simulation-based inference. In *Journal of Physics: Conference Series, 948*(1), 12-33. IOP Publishing.

Schuchardt, A. M., & Schunn, C. D. (2016). Modeling scientific processes with mathematics equations enhances student qualitative conceptual understanding and quantitative problem solving. *Science Education, 100*(2), 290-320.

Schwab-McCoy, A. (2019). The State of Statistics Education Research in Client Disciplines: Themes and Trends Across the University. *Journal of Statistics Education, 27*(3), 253-264.

Shuman L. J., Besterfield-Sacre M., Bursic K. M., Vidic N., T.P. Yildirim, and N. Siewiorek (2012). "CCLI: Model Eliciting Activities", In *Proceedings of the 2012 American Society for Engineering Education Annual Conference,* San Antonio, TX.

Sinharay, S., Haberman, S.J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26*(4), 21-28.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477-481.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement, 50*(1), 99-104.

Sireci, S. G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*(1), 100-107.

Sireci, S. G., & Sukin, T. (2013). Test validity. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbooks in psychology®. APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (p. 61–84). American Psychological Association.

Spence, D. J., Bailey, B., & Sharp, J. L. (2017). The Impact of Student-Directed Projects in Introductory Statistics. *Statistics Education Research Journal, 16*(1), 240-261.

Stanhope, L., Ziegler, L., Haque, T., Le, L., Vinces, M., Davis, G. K., ... & Umbanhowar Jr, C. (2017). Development of a biological science quantitative reasoning exam (BioSQuaRE). *CBE—Life Sciences Education, 16*(4), ar66.

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC medical research methodology, 8*(1), 45.

Tintle, N., Clark, J., Fischer, K., Chance, B., Cobb, G., Roy, S., ... & VanderStoep, J. (2018). Assessing the association between precourse metrics of student preparation and student performance in introductory statistics: Results from early data on simulation-based inference vs. nonsimulation-based inference. *Journal of Statistics Education, 26*(2), 103-109.

Tintle, N. L., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., … Vanderstoep, J. (2014). Quantitative evidence for the use of simulation and randomization in the introductory statistics course. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics,* Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

Tintle, N. L., Topliff, K., VanderStoep, J., Holmes, V. L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal, 11*(1), 21-40.

Tintle, N., VanderStoep, J., Holmes, V. L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education, 19*(1), 1-25.

Tintle, N., & VanderStoep, J. (2018). Development of a tool to assess students' conceptual understanding in introductory statistics. In *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10),* Kyoto, Japan: International Statistical Institute.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., ... & Hempel, S. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Annals of internal medicine*, *169*(7), 467-473.

Vidic, N. S., Ozaltin, N.O., Besterfield-Sacre, M., Shuman, L., (2014, June). Model Eliciting Activities motivated problem solving process: solution path analysis. In *Proceedings of the 121 ASEE Annual Conference*. Indianapolis, IN.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Wang, P., Palocsay, S. W., Shi, J., & White, M. M. (2018). Examining Undergraduate Students' Attitudes toward Business Statistics in the United States and China. *Decision Sciences Journal of Innovative Education, 16*(3), 197-216.

Ziegler, L., & Garfield, J. (2018). Developing a statistical literacy assessment for the modern introductory statistics course. *Statistics Education Research Journal, 17*(2), 161-178.

Zimmerman, W. A., Kang, H. B., Kim, K., Gao, M., Johnson, G., Clariana, R., & Zhang, F. (2018). Computer-automated approach for scoring short essays in an introductory statistics course. *Journal of Statistics Education, 26*(1), 40-47.

Zonnefeld, V. L. (2015). Mindsets, attitudes, and achievement in undergraduate statistics courses (Doctoral dissertation, University of South Dakota).

V.N. VIMAL RAO
56 E River Road Rm 250
Minneapolis, MN, 55455

Dissertation Abstract – VN. Vimal Rao

With scholars decrying a bastardization of the classical statistical testing procedure that has promulgated throughout common practice, many are turning away from statistical tests and their misunderstood null hypotheses and *p*-values. The issue at the heart of this matter is that the logic of statistical tests is confusing, the manner in which statisticians formulate their expectations as a probability distribution is poorly understood, and more generally, reasoning about distributions is difficult.

Over the years, statistics educators have tried many ways to improve their teaching and communication of statistical theory and practice. Starting in the 1980s, they began to utilize simulation. The statistics education community generally believed that this pedagogy was more apt to developing students' conceptual understanding of statistics, by placing the logic of statistical inference at the core of instruction and eschewing units on probability that was challenging for many students.

Considering the case of graduate students, as researchers applying statistical methods or practitioners interpreting statistical results, it's clear that they need to understand the logic of statistical inference. As science is fundamentally about theory generation and theory testing, it's also clear that they need to be fluent in at least one method of statistical testing. Might simulation based inference (SBI) curricula be able to support graduate students' development of an understanding of the core logic of statistical testing? How do graduate students, who have completed an SBI course, think while conducting statistical tests?

To answer these questions, a multi-model multiple descriptive case study of graduate students was conducted. Six graduate students in the educational sciences were recruited to complete the study approximately 7 months after they had completed an introductory SBI statistics course. Data sources included audio, video, and gaze recordings, analytic memos generated by the researcher, as well as written artifacts generated by the participants. Participants generated concept maps, conducted statistical tests, interpreted statistical results from tests, and participated in a video-cued retrospective interview. Data analysis was conducted through an interpretivist epistemological stance and employed the constant comparative method to identify relevant moments across all data artifacts.

Preliminary analyses suggest that students' statistical reasoning (i.e., their understanding of why a specific statistical procedure should be conducted and what can be learned from it) was generally quite good, although there were many gaps in their statistical thinking (i.e., their ability to determine which statistical procedures should be done at what time and how to enact those procedures). Students generally struggled in thinking about null hypotheses and the probability distributions they specified, as well as with thinking about *p*-values, instead focusing on point and interval estimates for statistics of interest. In particular, students struggled to contextualize the null hypothesis and *p*-value, but readily saw the connection between point and interval estimates and the original problem context. When students did consider *p*-values, they seemed to only vaguely remember that "< .05 is something".

This study is one of the first to examine graduate students' statistical thinking several months after the completion of an SBI introductory course. It is expected that students will forget many of the details taught. What they did remember, high level reasoning and a focus on variability through the examination of point and interval estimates, suggests that statistics instructors might anchor instruction about statistical inference and tests to descriptive statistics and their interpretation and contextualization. These results allow us to evaluate the potential benefits of an SBI curriculum for these students, identify conceptual difficulties that can be addressed with pedagogical reform, and inform reformation in the practice of statistics to address extant controversies and crises.