Extrapolation Under Caricatured Representations

Daniel C. Silliman (dsillim1@binghamton.edu)
Kenneth J. Kurtz (kkurtz@binghamton.edu)

Department of Psychology, Binghamton University (SUNY)
Binghamton, NY 13902 USA

Abstract

Research on contrastive category learning has revealed a robust tendency for learners to develop *caricaturized representations* (elsewhere: *ideals* or *extreme points*) to support successful discriminative classification. These representations are defined by extreme values on some task-relevant dimension and are often indicated as highly representative of their categories. Work in this area has elaborated the task constraints and contexts necessary for these representations to emerge, but little research has scrutinized whether caricatured representations extend beyond a category's known range of feature values. To these ends, across two experiments, we investigated whether the most representative items for a category can extend beyond the training set. Data from pairwise typicality comparisons following learning suggests that caricatured categories may be supported by representations that extend past the feature range present in training. The findings are better explained by certain representational frameworks (e.g., adaptive reference points, boundaries) than others (e.g., exemplars, clusters).

Keywords: categories; learning; caricatures; ideals; generalization; typicality; representation

Introduction

Since the highly influential investigations into the graded structure of categories (Rosch & Mervis, 1975), various researchers have sought to further explore not only the moderating and mediating factors that define the representative locus of a category, but also how perceived exemplar typicality in turn affects category use. Whereas a great deal of that follow-up work assumed typicality to be a function of taxonomic (i.e., feature-based) similarity to some mean or modal category representation, a study conducted by Barsalou (1985) revealed that typicality can also be understood in terms of the degree to which an exemplar embodies some dimension that furthers the ends of a goal-derived category—e.g., *caloric value* for *things not to eat on a diet*. Exemplars that realize the extreme values along those dimensions were considered *idealized* and often perceived as highly typical (for a dissenting view, see Kim & Murphy, 2011).

Though Barsalou's (1985) original paper discussed idealness in relation to natural language, goal-derived categories, subsequent work has since expanded the construct to apply to taxonomically defined categories as well (Davis & Love, 2010; Levering & Kurtz, 2006). Insofar as a learner's goal is to optimize performance on some classification task, any feature dimension with high diagnostic value can be construed as furthering the ends of that goal. It follows that exemplars with feature values instantiated on the higher or lower range of that dimension—the category *caricatures*—are often seen as both highly idealized and highly typical (Ameel & Storms, 2006; Davis & Poldrack, 2013).

Much as with the *peak-shift* effect in single-dimension stimuli (Hanson, 1959), whether caricatured representations develop is contingent on the presence of a neighboring category structure (Goldstone, 1996)—specifically, whether context/task emphasizes that the concepts are interrelated or isolated. Further, the exact relationship of the categories has been shown to control whether the category representations are more aligned with prototypes or caricatures. For example, Levering and Kurtz (2006) demonstrated that the typicality distribution over a category's members shifts depending on whether its feature values occupy one end of the known range of values (i.e., learning with two unidimensional categories) or the middle of the known range (i.e., learning with three unidimensional categories).

Despite the literature detailing the enabling conditions of caricature effects, comparatively little investigation focuses on the mechanisms and representations that explain them. Davis and Love (2010) have proposed that error-driven learning can explain participants' preference for idealized exemplars. Their participants were tasked with classifying exemplars from two or more categories, with each category always contrasting with another on a single dimension. After learning, when asked to generate an average category member, responses skewed from the prototype in the direction away from the contrast category. The positions of the generated stimuli could be simulated using a prototype model employed in Sakamoto, Jones, and Love (2008) that allowed the prototype's location to move according to error,

rather than remaining fixed. The model operates via a singular representation that maximizes distance from the contrast category while minimizing distance to its own category members. Assuming the two categories are dissimilar enough to begin with, the representation would continue to resemble a classic prototype, as the model no longer must contend with maximizing distance between contrasting categories. This tracks with the earlier work by Goldstone (1996) suggesting that caricature effects emerge from two concepts being emphasized as interrelated.

We refer to models that learn localist representations as falling under an *adaptive reference point* framework. Allowing the category representations—be they singular (Sakamoto et al., 2008) or distributed (Jones & Love, 2006; Kurtz & Silliman, 2019)—to adapt to task constraints and feedback not only provides a means for models to account for caricature effects, but also yields novel, behavioral predictions. Consider the scenario where two categories overlap such that the optimal representation location(s) exceeds the known range of feature values. Under these circumstances, the locus of category representation lies beyond the training set. Consequently, we could expect that the most typical items are not any of the experienced items but rather are extrapolated generalization items. Such an outcome would be difficult to reconcile with exemplar and cluster-based model frameworks which, while able to generalize to extrapolated items, may not view those items as more typical than the trained items. This is because the locus of representation for traditional reference point accounts always lies within the known range of feature values (i.e., either the items themselves or some reduced representation).

To our knowledge, none of the extant caricature literature has explored the typicality of extrapolated items under caricatured representations—most choosing instead to focus on enhancements or benefits afforded to ideal, known members. The study most germane to our conjecture comes from Nosofsky (1991), who manipulated the *presence of extreme-valued caricatures* along with the *frequency of instantiation for regular and extreme exemplars* in a classification task. After learning, participants engaged in a two-alternative forced-choice typicality task, where they had to choose the more representative item from a pair of the same category. The relevant finding concerns the condition where participants were not exposed to the extreme caricatures during learning, and the training exemplars with the end-range feature values were seen more frequently than others. When comparing these two types of items during the typicality preference task, participants favored the non-extreme, previously seen items. *Prima facie*, these findings do not suggest that extrapolated generalization items would be seen as more typical. However, Nosofsky (1991) was pitting frequency of instantiation against idealization, where the former has also been shown to contribute to category typicality (Barsalou, 1985). Nosofsky (1991) did not show that training exemplars with baseline frequency were favored over extrapolated caricature items.

The aim of the present study is to further this line of investigation and determine whether the locus of category representation can extend beyond known feature values. Concretely, this could be evidenced if participants prefer a test item with feature values beyond the range observed in training (extrapolated) as more typical than a test item with feature values within the observed range (interpolated). To test the above prediction, we conducted two experiments designed to encourage extreme caricature representations during a classification learning phase and probe said representations in subsequent generalization and 2-alternative forced-choice typicality preference (hereafter 2AFC) phases.

Experiment 1

The purpose of this experiment was to test whether caricatured representations could result in typicality preferences favoring extrapolated generalization items over interpolated generalization items. Such evidence would disfavor traditional reference point accounts that presume representation is fixed to experienced items or averages of such items. Toward testing this prediction, participants engaged in three phases: 1) a classification learning task with two, symmetrical, continuous-valued, overlapping categories that shared a diagnostic dimension, 2) a generalization test phase with interpolated items, extrapolated items, and several items from the classification phase, and 3) a 2AFC typicality phase featuring the items from the generalization phase. The 2AFC was included because it allows us to directly compare how representative items are within a category, rather than between possible categories (as with generalization).

Based on the number of highly confusable, overlapping items, we anticipated that final block accuracy would be around 60-80% for the initial classification phase—the upper bounds being contingent upon whether participants can successfully commit difficult items to memory. This range of accuracy is not an issue for this experiment, as we only require the participants to perform well enough to develop caricatured representations (as is the intent of this phase). We further predicted that participants should perform near ceiling for the more ideal items on the generalization phase—contingent upon their learning the diagnostic dimension from the preceding phase. Differences between the two critical test items, interpolated and extrapolated (see Figure 1), should be negligible. Items with more extreme values should be confidently categorized correctly, and as such, will likely be susceptible to ceiling effects. The generalization phase is intended primarily as a manipulation check of the category structure and task.

Regarding the 2AFC phase, we predicted that participants should significantly favor the extrapolated generalization item to the interpolated generalization item. Several other pairwise comparisons from this phase were also analyzed but are tangential to our primary question and were therefore exploratory in nature.

Methods

Participants The precedent for this design (Nosofsky, 1991) used 50 participants per condition. In anticipation of stringent screening criteria following online data collection, we collected roughly twice this amount. 92 psychology undergraduate students from Binghamton University participated in the study in exchange for partial course credit. Participants who reported a personal technical issue (n=3), who responded in a way indicating that they did not understand the instructions properly (n=4), or who self-reported multi-tasking during the study (n=13), were dropped from the dataset. These screening criteria left 72 participants (~78% of original data) for analysis.

Materials and Design The stimuli consisted of squares that varied continuously on dimensions of size (in cm) and shading (grayscale values). A previous study by Conaway and Kurtz (2017) demonstrated that these two dimensions were equally salient to participants. From these dimension, two overlapping categories (10 stimuli each) were constructed for the classification phase (see Figure 1). The semi-diagnostic dimension that delineated the categories was shading, though the dimension lost utility near the overlapping items (comprising ~40% of each category).

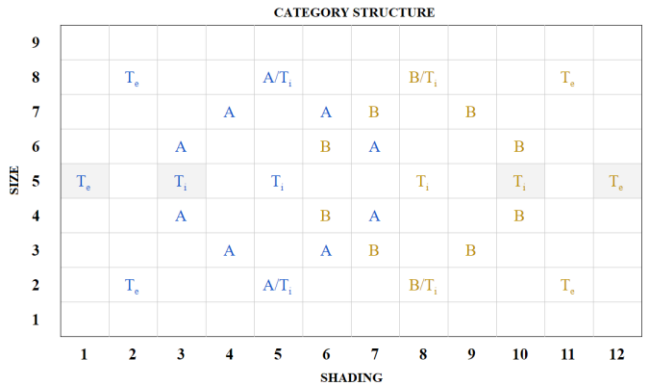

Figure 1: Categories used for Experiment 1. Category (A or B) indicated by colored letters. *T*s indicate items used in generalization/2AFC. Subscript *e* denotes extrapolated items, *i* denotes interpolated items. Grey backgrounds denote prediction-critical within-category pairs. Axis values are arbitrary and do not represent actual numerical values used for stimulus construction.

Though multiple interpolated and extrapolated items were used in each phase, the two critical pairs always consisted of the furthest extrapolated item (i.e., most extreme along the diagnostic dimension) and the furthest interpolated item. A programming error resulted in training item "10-6" being omitted from the learning phase, while training item "9-7" was seen twice as often. Though inconvenient, neither item was intended to be involved in the two test phases, and the nature of the error does not disrupt the formation of caricatured representations.

Procedure All data collection occurred online via a lab-hosted web server due to COVID-19 restrictions on in-person research. Following informed consent, participants immediately began the classification phase. A general instruction screen presented participants with a brief description of the task and cover story. Participants were told that the squares were the same hieroglyphic letter from two related but distinct cultures, and that the hieroglyphs varied within a culture (as in differences with handwriting/style), but the variation between cultures was greater. The category labels were the cultures themselves (located in the SOUTH and WEST). Participants were also told that they would be receiving corrective feedback, and that while they would have to guess at first, that they will eventually come to understand what defines each category.

On each classification trial, participants were presented with a single stimulus centered onscreen. Two onscreen response buttons labeled SOUTH and WEST were positioned below the stimulus. Trial-wise instructions at the top of the screen queried which group the hieroglyph belonged to. Participants were given unlimited time to make a response. Upon selecting a response, feedback was presented while the stimulus was still onscreen either notifying the participant that they had chosen correctly or incorrectly—always providing the correct category label. Feedback was kept onscreen until the participant clicked a continue button. There was no inter-trial interval (ITI) between feedback and the start of the next trial. Item order within a block was assigned randomly. This phase repeated for five blocks and a total of 100 trials of classification.

The generalization phase followed the completion of the classification phase. Participants were instructed at this point that they would now be tested on their knowledge of the two categories. They were told that they would be seeing a few old hieroglyphs and several new ones and that their task was much the same as before except that now they only had one chance to accurately classify a hieroglyph and would receive no feedback. Aside from the differences mentioned in the instructions, we now included an ITI of 200ms to prevent accidental double-clicks across trials. Following all 14 trials in the generalization phase, participants proceeded to the 2AFC phase.

At the start of the 2AFC phase participants were instructed that their task was to review two examples from the same category and indicate via mouse click which of the two examples was more representative of its category. A representative hieroglyph was defined as an example that embodied what they believed made a hieroglyph a good example of its category. An example was provided using the superordinate category of *birds*, and the subordinate categories of robins (good example) and ostriches (poor example). Participants were informed that there would be no feedback as there was no objectively 'correct' answer. On each trial two stimuli from the same category of the preceding phase were randomly positioned to either the left or right of the screen. A prompt at the top of the screen provided the

category label and asked participants to click on the more representative example. Participants were given unlimited time to make their choice. Every pairwise combination of within-category pairs from the generalization phase was used with the order randomized. After concluding the experiment, participants were given the opportunity to complete an exit survey that was the basis for the screening described above.

Results and Discussion

All analyses were conducted in the R programming environment. In accord with initial predictions, final block classification accuracy was ~70% (*SD* = 0.45). Participants who scored 50% (chance) or less in the final block were excluded from further analyses. This resulted in eight participants being dropped from the dataset (~11% of the data following the screening outlined earlier) which left 65 participants for further analysis.

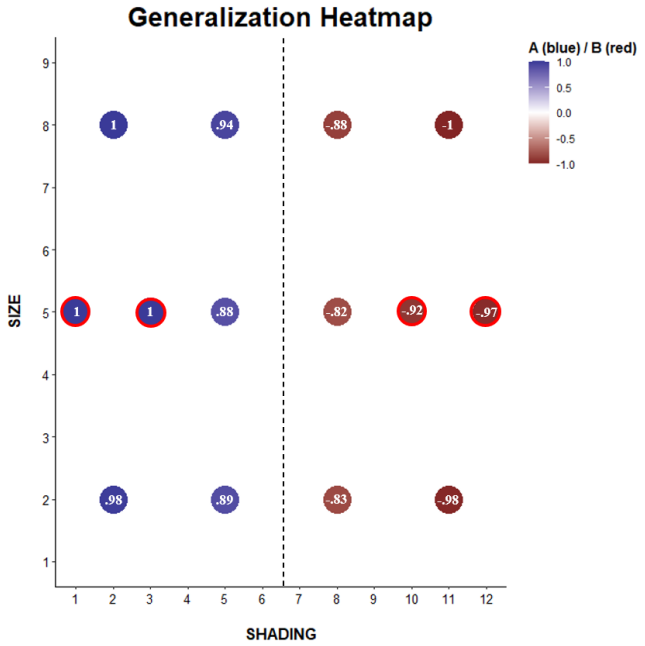

Figure 2: Heatmap of generalization for Experiment 1. Category choices recoded as +1/-1 and averaged. Critical items are outlined in red.

In accord with our predictions regarding generalization, both the critical interpolated and extrapolated item were correctly categorized with high accuracy (see Figure 2). Two generalized linear models (one per category) were built to determine if the furthest extrapolated items were generalized accurately significantly more often than all other items. An ICC revealed ~10% of total variance was attributable to between-subjects variance, suggesting a random effects structure was unnecessary. Generalization accuracy was predicted from stimulus position. For the model on *Category A* items (red items of Figure 2), no significant differences were found (all *ps* > .993). For the *Category B* model, the extrapolated item was found to be accurately categorized significantly more often than items "8-2" (β = -1.859, *SE* = 0.790, *p* = .019) and "8-5" (β =-1.965, *SE* = 0.786, *p* = .013)—all other *p*s > .067. The asymmetry in model results may be tied to the missing item during training—"10-6". Though unexpected, this finding does not affect our conclusions, as we were mainly concerned with the furthest extrapolated and interpolated items.

Our primary prediction for the typicality phase was that the most extreme-valued extrapolated items ("1-5" and "12-5") would be selected as 'more representative' significantly more often than the most extreme-valued interpolated items ("3-5" and "10-5") when the two were directly compared. Because the preference proportions collapse across multiple comparisons per participant (see Figure 3 caption), analyses on those proportions would violate assumptions of independence inherent to most tests. Instead, we examined only the subset of comparisons that include the two critical items—resulting in a single response per category per participant. Responses favoring the interpolated item were re-coded to be negative. A binomial test was conducted to detect significant difference from chance. This process was repeated for each category separately.

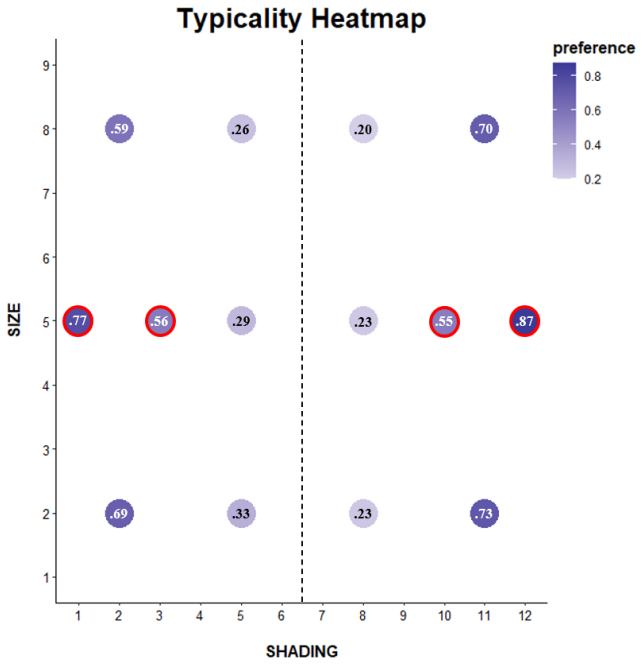

Figure 3: Heatmap for the 2AFC typicality task in Experiment 1. The sum for each item was divided over the total number of comparisons. The purple border denotes critical items.

For *Category A* items, the analysis revealed that the extrapolated test item was selected significantly more often (54 to 11, *p* = .001). The same pattern of significance is found for the *Category B* items (55 to 10, *p* = .001). Cohen's *g* values for the differences are 0.33 and 0.361, respectively (both large). A *post-hoc* power analysis was conducted to determine the adequacy of our sample size. Assuming the lowest observed preference (54/65 = .83) as our alternative

hypothesis, at N=65 participants, we would have power > .99 to reject the null. Exploratory analyses for *Category A* revealed that 49 participants preferred item "1-5" (the most extreme item) to its upper flanker, "2-8" (p < .001), and 45 participants preferred item "1-5" to its lower flanker, "2-2" (p = .002). When comparing *Category B* items 54 participants preferred "12-5" to "11-2" (p < .001), and 52 participants preferred it to item "11-8" (p < .001).

These data suggest that the locus of caricature representation for a category may extend beyond the known range of exemplars. These findings are incompatible with traditional reference point frameworks (e.g., clusters, exemplars) and suggest that a different framework may be warranted. Although these data are better explained by an adaptive reference points framework, they are equally well explained by category boundary models, wherein increasing psychological distance from an optimal decision bound can be interpreted as increasing confidence in a category decision (Ashby et al., 1998). We address these competing explanations in the following experiment.

Experiment 2

The aim of Experiment 2 was to better dissociate the competing explanations in Experiment 1 as well as to provide a partial replication. Much of the design remained the same, with the critical difference being in the appearance of the most extreme extrapolation items. These items were now made more deliberately extreme so that they were considerably different from the end-range training items. Such extreme items should always be favored as more representative if the prior results arise from the mechanisms of a decision bound framework—recall that further psychological distance from a bound should always translate to greater confidence in category membership. If, however, participants view the items as too different from the training set, and reject them in favor of the interpolated item, such behavior would be better explained by adaptive reference points. This is because adaptive reference points are still constrained by the need to minimize within-category distance. Consequently, typicality would begin to fall off after some distance from the reference points. In addition to the aforementioned changes, we also omit the generalization phase as it is less informative for the present experiment. Further, we also swap the dimensions from Experiment 1 (see Figure 1), such that size is now the diagnostic dimension. Doing so allowed for greater range in the extreme extrapolated items.

Methods

Participants A total of 106 participants from Binghamton University were run in this experiment. After dropping participants for self-reporting multi-tasking (n=9), and another for self-reporting display issues, 96 participants (~91% of the original data) were left for analysis.

Materials and Design The domain was the same, however, the diagnostic dimension was now size. The smallest square seen during classification was 3.52 cm, while the largest square seen during classification was 6.58 cm. The most extreme extrapolations were 0.5 cm and 16 cm for small and large, respectively. Both values were made as extreme as possible while still limited by factors of visibility (small) and the average monitor resolution of our participants (large). This decision resulted in a minor asymmetry between the differences of the largest and smallest seen and unseen items.

Procedure The classification phase was largely the same, however, the cover story was changed in light of a small minority of participants (n=4) reporting being confused about the prior cover story; the squares were now characterized as sheet metal produced by two separate companies. The same labels (SOUTH and WEST) were kept.

Results and Discussion

Final block accuracy for the classification task was 65% (*SD* = 0.47). After dropping participants who did not meet the 50% learning criteria (~23% of the data), 74 participants remained for the analyses on the 2AFC task.

The aim of the 2AFC typicality phase was to determine if participants preferred the extreme caricatures—thereby supporting category boundaries—or rejected them, thereby supporting adaptive reference points. Inspection of the 2AFC item preference provides mixed support for both accounts (see Figure 4). There appears to be an asymmetry in preference for the extremely small caricature and the extremely large caricature. This asymmetry is not explained by the caricatures' physical similarity to the nearest seen exemplar. Were this the case, we would expect the small caricature to be favored more so than the large caricature, as it is more similar (physically) to the training set. There is a possibility that 'bigness' defined one category more than 'smallness' defined the other, but the cause for this systematic bias is equally opaque.

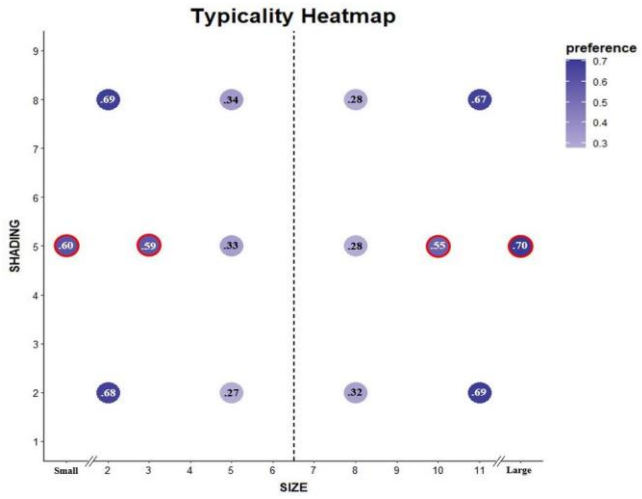

Figure 4: Typicality preferences for Experiment 2. For parity with the earlier figure, the same number of columns

are retained, however, the end columns are not one unit's difference from 2 and 11—differing rather by several units.

Our primary analyses concern the two critical within-category pairs (see Figure 4). Consistent with the category boundary models, the binomial tests revealed that participants still significantly preferred the critical extrapolated item to the critical interpolated item for both categories (*Category A*: 48 to 26, $p = .014$; *Category B*: 54 to 20, $p < .001$). Notably, for both categories, participants' preference for the extrapolated item is reduced from that seen in Experiment 1, though still greater than chance. As in Experiment 1, we also conducted exploratory analyses comparing the most extreme-valued extrapolations with its nearest extrapolated flankers ("2-8" and "2-2" for *Category A*, "11-8" and "11-2" for *Category B*). Curiously, when comparing the small item with its flankers, it is not significantly preferred to either "2-8" (41 to 33, $p = .146$) or "2-2" (44 to 30, $p = .13$). This is not true for the large item, which is still preferred to its flankers—items "11-8" (53 to 21, $p < .001$) and "11-2" (51 to 23, $p = .001$). Contrary to the main finding, the observed preference for the flanker extrapolated items ("2-8", "2-2") over the most extreme-valued extrapolated item (small), is more consistent with the adaptive reference points framework (Davis & Love, 2010, Kurtz & Silliman, 2019).

The asymmetry in overall preference for the extreme caricatures (across multiple comparisons), while not intended, is also better explained by an adaptive reference points framework. Assuming a category boundary explanation, both extreme caricatures should be equally preferred to the remaining test items, as each is respectively the furthest from the ostensible boundary demarcating its category. While a boundary framework cannot admit of this asymmetry, adaptive reference points can be positioned independently of each other, adopting an arbitrary constellation of positions if it best suits the task and stimuli.

General Discussion

Across two experiments we have found novel and robust evidence of caricature representations that extend beyond the known training set. To date, previous investigations of caricatures have been limited to demonstrating advantages for ideal training exemplars and generated stimuli that deviated from the prototype position. The present results are the first indication that caricatured categories can result in the most representative items lying beyond even the most extreme observed items.

There are a few caveats to the present findings that warrant discussion. First, the procedure of screening participants who did not achieve greater than 50% accuracy in the final block has the potential to invite selection effects. This procedure is not uncommon in the categorization literature, as predictions concerning possible outcomes of test phases are predicated on the assumption that the participant has learned something about the categories. There is a possibility, though, that the difficulty of our task filters for only the type of learner who would respond in the observed manner for the latter test phases. It is entirely possible that a different design that is easier to learn may result in a different pattern of findings due simply to differences in the pool of participants being analyzed. While our procedure is vulnerable to this risk, it is a necessary risk, as the alternative of analyzing all non-learners is equally undesirable. To mitigate this risk, we set our filter to a very achievable number—greater than 50%.

A related concern is the difference in the number of participants filtered after classification between Experiments 1 (~11%) and 2 (~23%). While some portion of the difference is likely owed to noise, it is more likely that swapping the dimensions along the x/y axes played a role. To reiterate from the Experiment 1 methods, this learning domain was selected because it had previously been normed such that each dimension was found to be equally salient to participants (Conaway & Kurtz, 2017). It is possible, however, that participants are treating the dimensions differently at lower vs. higher values—e.g., a 1-units difference at the high end of the range is perceived as much greater than a 1-units difference at the low end. This is likely not the case for Experiment 2, given that prior psychophysics studies suggest that the function relating perceived changes to line length and objective changes to line length is linear (Stevens, 1957). It is more a concern for Experiment 1 (lightness of gray squares), where the exponent of the power law relating perceived to actual changes is closer to 1.2 (Stevens, 1957). This may explain why accuracy was higher for Experiment 1. If the representations of one category are much further away in psychological space from the other category, that category would be easier to learn, as its items are less perceptually confusable with the contrast category.

While not ideal, the potential differences in how participants perceived the diagnostic dimensions is only a problem for our conclusions if said differences impact caricature representation. Under both theories of caricature effects—category boundaries and adaptive reference points—this is not a concern. In the case of the former, the most representative item is simply that item which is furthest from the bounds. Even if one category is much more psychologically distant from the other category, there will still be a "furthest item". Likewise, for adaptive reference points, we can assume that the reference points themselves exist on the same scale as the stimulus representations. Therefore, as one dimension expands at higher physical values, so too will the reference points move with them. In either case, we would expect the same predictions for both Experiment 1 and 2.

Regarding theoretical implications of this work, the findings are most compatible with either a category boundary or adaptive references points framework. Traditional reference point accounts are principally unable to account for these data, though proper simulations will be needed for a more definitive conclusion on this matter. For example, cluster or exemplar accounts may explain the 2AFC data if they assume a Gaussian similarity gradient, rather than an exponential similarity gradient (Nosofsky, 1991)—though it

remains to be seen how this will impact model fits of the classification data. Likewise, comparing simulations of category boundary and adaptive reference point models will be equally informative as to the results of Experiment 2.

We note here that although we have to this point been discussing the competing frameworks as though they were mutually exclusive, it is entirely possible that participants are leveraging multiple systems of categorization to complete our tasks. Though assuming two or more distinct architectures is less parsimonious, there is growing evidence that these distinct architectures may be instantiated in functionally distinct parts of the brain (Ashby et al., 1998; Bowman, Iwashita, & Zeithamova, 2020). Under this combined framework, participants could be averaging over the outputs of these systems or even alternating between them on a per-trial basis or contingent on the nature of the task. Our data may even reflect a mixture of participants, some of whom prefer to use boundaries, while others preference adaptive reference points or exemplars/clusters. Though these possibilities are merely speculative for now, future theoretical and empirical work may shed light on their respective likelihoods.

Assuming one framework or the other, however, the data are inconclusive as to whether adaptive reference points or category boundaries undergird the effects. While the results of the critical pair comparisons in Experiment 2 are more consistent with category boundaries, the asymmetry in preferences is better explained by adaptive reference points. We also note reduced preference for extreme caricatures relative to the critical interpolated item in Experiment 2. It may simply be that the extreme extrapolations were not distinct enough from the category distributions to conclusively show that the majority of participants reject them.

Determining which of these two accounts is more likely to explain caricature effects is critical to the endeavor of understanding category representation. To the extent that graded structure is a critical component of categorization, and to the extent that caricature effects represent a distinct form of graded structure, one can argue that our understanding of categorization will remain incomplete so long as our present theories are unable to fully and conclusively account for this phenomena. It is for these reasons that future empirical work on caricature effects should prioritize determining which theory of category representation (or combination thereof) best explains the data.

References

Ameel, E., & Storms, G. (2006). From prototypes to caricatures: Geometrical models for concept typicality. *Journal of Memory and Language*, *55*(3), 402-421.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological review*, *105*(3), 442.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 629.

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *Elife*, *9*, e59360.

Conaway, N., & Kurtz, K. J. (2017). Similar to the category, but not the exemplars: A study of generalization. *Psychonomic bulletin & review*, *24*(4), 1312-1323.

Davis, T., & Love, B. C. (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science*, *21*(2), 234-242.

Davis, T., & Poldrack, R. A. (2013). Measuring neural representations with fMRI: practices and pitfalls. *Annals of the New York Academy of Sciences*, *1296*(1), 108-134.

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & cognition*, *24*(5), 608-628.

Hanson, H. M. (1959). Effects of discrimination training on stimulus generalization. *Journal of experimental psychology*, *58*(5), 321.

Jones, M., & Love, B. C. (2006). The emergence of multiple learning systems. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 28, No. 28).

Kim, S., & Murphy, G. L. (2011). Ideals and category typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1092.

Kurtz, K. J., & Silliman, D. C. (2019). Warning: The exemplars in your category representation may not be the ones experienced during learning. In A. Goel, C. Seifert, & C. Freska (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society (pp.* 56-57*)*. Montreal, QB: Cognitive Science Society.

Levering, K., & Kurtz, K. J. (2006). The influence of learning to distinguish categories on graded structure. In *Proceedings of the 28th annual conference of the Cognitive Science Society* (pp. 26-29). Austin, TX: Cognitive Science Society.

Nosofsky, R. M. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition*, *19*(2), 131-150.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573-605.

Sakamoto, Y., Jones, M., & Love, B. C. (2008). Putting the psychology back into psychological models: Mechanistic versus rational approaches. *Memory & Cognition*, *36*(6), 1057-1065.

Stevens, S. S. (1957). On the psychophysical law. *Psychological review*, *64*(3), 153.