Contrast Categories Elicit Illusory Correlations When Learning Social Groups

Kimery Reed Levering (kimery.levering@marist.edu)
Brittany Mross (Brittany.mross1@marist.edu)
Natalie Bills (natalie.bills1@marist.edu)
Mallory Cannon (mjc0113@auburn.edu)
Jacqueline Cassano (jacqueline.cassano1@marist.edu)

Department of Psychology, Marist College
3399 North Road, Poughkeepsie, NY 12601 USA

Abstract

Assigning category labels to examples varying along a continuous dimension exaggerates perceived differences between members on opposite sides of category boundaries. Using social categories, we investigated how contrast may guide representation of not only features that differentiate between groups but also features that are neither diagnostic nor correlated across trained examples. In a classification task, participants learned which residence hall to correctly assign students varying along three psychological traits (academic, athletic, social). The same target category was learned alongside one of two contrast categories with either higher or lower values along one diagnostic dimension. After learning, participants provided estimates of average trait values for each dorm. Predicted contrast effects were found along diagnostic dimensions but contrast also influenced memory for non-diagnostic and uncorrelated dimensions, presumably based on assumptions about general co-occurrence of features. These findings have implications for how stereotypes are learned and applied and how illusory correlations are perpetuated.

Keywords: categories, concepts, learning, contrast, stereotypes, illusory correlation, ideals

Introduction

It has long been demonstrated that associating category labels with exemplars varying along a continuous dimension increases perceived differences between members that straddle the category boundary (Goldstone, 1996; Goldstone, Steyvers, & Rogosky, 2003; Harnad, 1987; Palmeri & Nosofsky, 2001; Tajfel & Wilkes, 1963). In a classic experiment, Tajfel & Wilkes (1963) presented 8 lines varying in length and asked participants to judge the length of each. When the four shortest lines were labeled with a different category name (A) than the other four lines (B), the longest A line and the shortest B line were judged to be more different from each other in length than they were in reality. Further research has shown that after learning, category members furthest from a category boundary are classified more accurately and processed faster (Beale & Keil, 1995; Goldstone, 1996).

Beyond demonstrated effects on ease of perceptual differentiation and classification, there is evidence that the nature of a co-learned contrast category can affect category representations themselves. Manipulating what a target category is learned against has been shown to affect ratings of how typical members are of the target category (e.g., Davis & Love, 2010; Levering & Kurtz, 2010) and also memory for perceived feature averages across all members of a category. For example, Davis & Love (2010) exposed participants to examples of categories varying along two dimensions (either energy sources varying in cost and level of pollution or supporters of political candidates varying in age and income level). Category membership reflected a unidimensional rule along either one or the other dimensions (or both). When later asked to adjust the amount of pollution or cost to match their memory of the average value, participants remembered averages as further from contrast categories than they actually were. Using fMRI, Davis & Poldrack (2014) later found that extreme examples were not only thought of as more typical but were associated with patterns of neural firing in the temporal and occipital cortex that most closely matched other members of their category.

In essence, the representation of a category is a caricature in relation to opposing categories, rather than reflecting the central tendency (average) of its members. These kinds of effects have been referred to as learned categorical perception, accentuation, or contrast effects and have been found not only with basic perceptual categories like line lengths (Tajfel & Wilkes, 1963), shapes (Katz, 1963), and objects (e.g., Livingston, Andrews, and Harnard, 1998) but also categories of people, based on both physical characteristics like faces (e.g., Beale & Keil, 1995; Goldstone, 1996), and psychological characteristics like attitudes (e.g., Eiser, 1971) and traits (Krueger & Rothbart, 1990).

Idealized representations of categories are believed to arise through the consideration of goals along certain dimensions in goal-derived categories like "things to eat on a diet" (Barsalou, 1985) but also through the act itself of differentiating or comparing between categories. For example, Goldstone (1996) found greater focus on diagnostic dimensions in categories that are highly interrelated with other categories (as opposed to isolated) and Goldstone, Steyvers, and Rogosky (2003) found that as categories were moved apart from each other (larger differences between

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

contiguous examples from different categories), contrast effects can be mitigated.

Findings from Davis and Love (2010) strongly suggest that contrast categories determine category ideals only along dimensions highlighted by the particular distinction learners are asked to make. Specifically, they found that only contrasted features on which a specific contrast was made during learning (e.g., cost of energy source) demonstrated a shift of representation away from co-learned categories. Levering & Kurtz (2015) found that after both supervised classification and supervised observational learning, regularities along dimensions not necessary for classification were not only not accentuated but were simply not as well learned. This could be due to a perceptual focusing on relevant features, at the expense of nondiagnostic ones. For example, Goldstone (1994) found acquired equivalence along irrelevant dimensions for separable dimensions, meaning that a category distinction that was based on a relevant feature (e.g., size) made differences along one that was not relevant (e.g., brightness) less discriminable in a later perceptual task. Category representations that include information outside what is relevant for determining category membership (such as within-category statistical regularities) has been demonstrated, but typically through learning tasks that do not focus the learner as much on discriminating categories, such as unsupervised learning (e.g., Kemler Nelson, 1984), inference learning (e.g., Chin-Parker & Ross, 2004), observational learning (e.g., Hoffman & Murphy, 2006), and using category members to make inductions or predictions (e.g., Minda & Ross, 2004).

The Role of Illusory Correlations

While informative, these studies looking at the effect of contrast on representations of nondiagnostic features (e.g., Davis and Love, 2010; Goldstone, 1994; Levering & Kurtz, 2015) used stimuli with features free of strong pre-existing associations. In the real world, we have theories about what features co-occur and these theories can guide representation of the features that define membership and also perhaps related features. Particularly in social domains, we often have (potentially incorrect) theories about what features co-occur, a phenomenon thought to lead to the maintenance of social stereotypes (Hamilton & Rose, 1980). For example, knowing that a group of people are athletic (relative to others) may lead us to believe that they also eat healthy, are more social, etc. despite a lack of demonstrated evidence of these traits.

Perceived relationships between features that are only weakly or not at all correlated (illusory correlations) was originally studied in the context of clinical diagnoses. In work that coined the term illusory correlations, Chapman & Chapman (1967) had participants view drawings of humans along with two contrived statements about the person who drew the figures. Despite there being no relationship between features of the drawings (e.g., large or emphasized head) and personality traits (e.g., intelligence), participants consistently reported perceiving them. As emphasized by Nisbett & Ross (1980), perception of inaccurate correlations are less about the raw data and more about preconceived theories about associations that "ought to exist".

The Current Study

In the current study, we establish a paradigm to investigate how contrast may guide representation of not only the features that define membership but also other features that are neither diagnostic nor correlated within a training set. We made the three following predictions: (1) Traditional contrast effects would be found using this paradigm with psychological dimensions in social categories. Specifically, a target category with central values would be represented differently depending on what it is learned alongside. (2) Illusory correlations would arise along non-diagnostic and uncorrelated feature dimensions, despite a lack of category (or cue) validity. Depending on what it is being learned against, we predict the same category will be viewed as possessing different trait values along non-contrasting dimensions.

Method

Participants

147 Marist College students participated in exchange for partial fulfillment of course credit. All participants had normal or corrected-to-normal vision.

Stimuli and Category Structure

Stimuli (see Figure 1) were images of student profile cards which included a school logo, redacted information (student name, student ID number, and student year), and values for three psychological dimensions: Academic, athletic, and social preference. The dimensions were described as "scores", and were explained to comprised of several factors each. Scores could take on values between 0 and 10, and were specified to two decimal places. Possible values were equidistant from each other, creating a three-dimensional feature space with 30 values along each dimension.

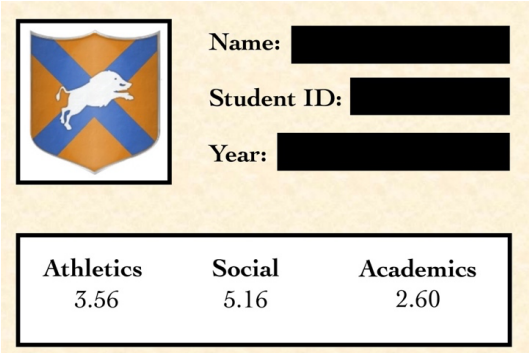

Figure 1: Example of a training item.

Category structure (see Table 1 for descriptive information about each) always followed a unidimensional rule, meaning that membership could be determined by looking at only one

of the three dimensions. Assignment of logical structure to feature values was fully counterbalanced so that for a third of participants the relevant dimension was academic, for a third of participants the relevant dimension was athletic, and for a third of participants the relevant dimension was social preference (see Table 2). Aside from the (counterbalanced) dimension associated with the unidimensional rule, the other two dimensions were always non-diagnostic. For example, participants assigned to a unidimensional rule along the academic dimension (groups 1 and 2) would learn that examples with higher academic scores are in one category and those with lower scores are in the other. For those participants, the social and athletic dimensions were non-diagnostic, as the means across items in both categories were 5.00 for those dimensions.

Table 1. Values of the three feature dimensions for each example

	CATEGORY A			**CATEGORY B**			**CATEGORY C**		
	F1	F2	F3	F1	F2	F3	F1	F2	F3
	9.00	**0.36**	9.32	9.64	**3.56**	9.32	9.00	**6.76**	9.32
	1.00	**0.68**	4.20	4.52	**3.88**	0.68	1.00	**7.08**	4.20
	8.04	**1.00**	0.68	0.36	**4.20**	8.36	8.04	**7.40**	0.68
	1.96	**1.32**	5.80	5.48	**4.52**	1.64	1.96	**7.72**	5.80
	7.08	**1.64**	8.68	8.36	**4.84**	7.40	7.08	**8.04**	8.68
	2.92	**1.96**	3.24	3.56	**5.16**	2.60	2.92	**8.36**	3.24
	6.12	**2.28**	1.32	1.64	**5.48**	6.44	6.12	**8.68**	1.32
	3.88	**2.60**	6.76	6.44	**5.80**	3.56	3.88	**9.00**	6.76
	5.16	**2.92**	7.72	7.40	**6.12**	5.48	5.16	**9.32**	7.72
	4.84	**3.24**	2.28	2.60	**6.44**	4.52	4.84	**9.64**	2.28
Mean	5.00	1.80	5.00	5.00	5.00	5.00	5.00	8.20	5.00
SD	2.62	0.97	3.10	3.01	0.97	2.91	2.62	0.97	3.10
	F1/F2	**F1/F3**	**F2/F3**	**F1/F2**	**F1/F3**	**F2/F3**	**F1/F2**	**F1/F3**	**F2/F3**
r	-0.15	0.15	-0.16	-0.18	0.15	-0.15	-0.15	0.15	-0.16

Note: Assignment of feature dimensions (social preference, athleticism, academic) to logical structure (F1, F2, F3) was counterbalanced. The diagnostic feature (F2) is in bold. All participants classified examples of category B (Mydro Hall), but half of participants learned to distinguish them from category A (lower values on F2) and half learned to distinguish from category C (higher values on F2).

A target category (Mydro Hall) was always made up of ten examples with central values along the contrasted dimension (B; M = 5.00) and participants were randomly assigned to learn a contrast category (Sorson Hall) that was made up of 10 examples with either smaller (A; M = 1.8) or larger (C; M = 8.2) values along that same dimension. To be clear, the target Mydro category (B) was identical for all participants, but the category could possess higher or lower values relative to the contrast category, depending on which condition they were assigned to. Values along non-contrasted dimensions were designed to have identical means across categories, making them not at all diagnostic for determining category membership. For example, those learning a unidimensional rule along the academic dimension, would not have improved performance by applying a rule along the athleticism dimension because members possessing both high and low values were members of each category. Correlations between feature values were minimal (none exceeded r = 0.18 for any category) and nearly identical between categories, meaning that relationships between features were also not diagnostic for membership. Relevant information was presented in the center of the card and the irrelevant features were on the left and right. Irrelevant feature order was counterbalanced.

Table 2. Description of the six conditions based on diagnostic dimension and contrast category.

		Diagnostic Dimension		
		Academic	**Athletic**	**Social**
Contrast Category	**A**	Group 1	Group 3	Group 5
	C	Group 2	Group 4	Group 6

Procedure

Training Phase After consenting to participate in the experiment, participants were given instructions asking them to imagine they were working in the residence office of a small liberal arts college. They were told that the previous person who worked in that position had suddenly quit without leaving information about how to assign residence halls. Their job was to learn the types of people placed in two male dormitory halls (Mydro Hall and Sorson Hall) so that they could later figure out where new students would be placed.

Participants were then explained the three pieces of information they would get about each student, including a description of the factors going into each.

Following self-reported understanding of the instructions, participants engaged in the training task. On each self-paced training trial, one of the 20 student profiles (10 from each category) was randomly selected and presented on the screen. Participants were instructed to decide which of two residence halls - Sorson Hall or Mydro Hall - the student had been assigned to by pressing one of two keys labeled with the names of the halls. The target category (B) was always labeled Mydro Hall and the contrast category was always labeled Sorson Hall. After responding, they were informed whether they were right or wrong and feedback about correct residence hall was given. Participants were trained on 4 blocks of classification learning, each consisting of all 20 student profiles, resulting in a total of 80 trials.

Testing Phase After training, participants were given two test phases (counterbalanced for order). In a feature inference test phase, they were asked to provide for each dorm an estimate of the average (mean) value along each of the dimensions, based on training. This is similar to the testing phase of Krueger & Rothbart (1990) and Love & Davis (2010). In a classification test phase, participants classified unlabeled examples without feedback. In addition to trained Mydro examples, participants were exposed to untrained examples that possessed even more extreme values along the relevant dimension. In other words, if they learned a distinction between A and B, they would then receive C members during test. These examples were included to investigate the impact of exposure to more extreme examples on perceived category boundary.

Results

Overall Learning

A Contrast Category (A, C) x Relevant Feature (Social, Academic, Athletic) x Learning Block (1, 2, 3, 4) mixed ANOVA performed on learning accuracy showed a significant effect of learning across blocks, $F(3,423) = 59.026$, $p < .001$, $\eta^2 = .106$, but no difference in learning based on contrast category or relevant dimension and no interaction between the factors, $ps > .05$.

Item-Level Contrast Effects

A separate analysis was conducted on learning performance to see if items were learned more easily when they were furthest from the contrast category, as seen elsewhere in the literature (e.g., Goldstone, et al., 1994). For each participant, we calculated the slope of the regression line based on their performance on items B01 through B10. A slope greater than zero indicates that learning performance increased as values along the contrasted dimension of B items increased. A Contrast Category (A, C) by Relevant Dimension (Social, Academic, Athletic) ANOVA on these slopes showed a significant effect of contrast category, $F(1,140) = 93.337$, $p < .001$, $\eta^2 = .394$. Specifically, Contrast A learners had significantly higher slopes (M = 4.470, SD = 4.727) than Contrast C learners ($M = -3.796$, $SD = 5.514$). As can be seen in Figure 2, this aggregate metric accurately reflected the performance of individual learners, as 86.111% of the Contrast A condition had positive slopes while 77.333% of the Contrast C condition had negative slopes. There was no main effect of relevant dimension nor an interaction between the factors, $ps > .05$.

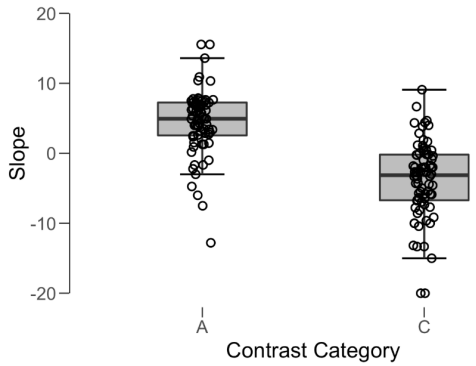

Figure 2: Distribution of training slopes across participants.

A three-way (Contrast Category: A, C; Relevant Feature: Social, Academic, Athletic; Post-Learning Exposure: Before Ratings, After Ratings) ANOVA was performed to compare relevant (diagnostic) feature estimates after learning for the target Mydro Hall (B) category depending on whether the contrast category contained lower or higher values and whether there was a difference depending on which dimension was relevant or whether they were exposed to more extreme unlabeled examples. As predicted, there was a main effect of contrast, $F(1,132) = 21.580$, $p < .001$, $\eta^2 = .138$[1]. When the contrast category possessed lower values than the target category, participants made higher estimates along relevant dimensions ($M = 5.801$, $SD = 1.95$) than when the contrast category possessed higher values ($M = 4.270$, $SD = 1.899$). There was no main effect of relevant category and no associated interaction between the factors, indicating that the contrast effects occurred similarly across relevant dimension conditions. There was also no main effect of post-learning exposure, indicating that expanding the range of examples had no measurable effect on category representation. This factor was omitted from further analyses.

To further investigate the extent of distortion from presented values, we computed difference scores based on how much each participant's perceived average differed from the actual average of the category, separately for each dimension. A positive difference score indicates that the

[1] Nine participants responded to at least one feature value question with a non-number and their data was excluded from analysis.

perceived average was higher than the actual average and a negative difference score indicates it was lower than the actual average. An independent-samples *t*-test showed that for relevant dimensions included in the unidimensional rule, difference scores for the Mydro category when compared to a higher Sorsen (M = -.730, SD = 1.899) was significantly lower than compared to a lower Sorsen (M = .801, SD = 1.947), $t(136) = 4.677$, $p < .001$, $d = .797$. Analyzed in separate one-sample t-tests, difference scores for the A condition was significantly lower than zero, $t(66) = 3.370$, $p = .001$, $d = .412$, and difference scores for the C condition was significantly higher than zero, $t(70) = 3.239$, $p = .002$.

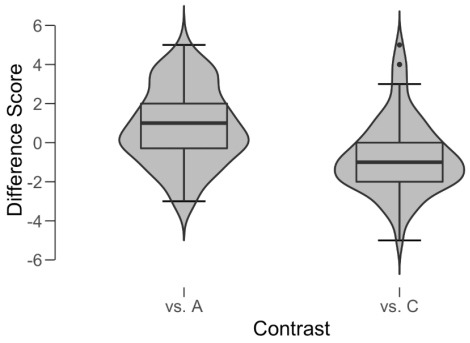

Figure 3: Distribution of difference scores (actual mean subtracted from reported mean) across participants.

Illusory Correlations

To examine the role of contrast on illusory correlations, we conducted a two-way ANOVA along non-contrasting dimensions comparing Contrast Category (A, C) and Relevant Feature (Social, Academic, Athletic). Despite the fact that these irrelevant dimensions had the same mean across both the target category (M = 5.00) and contrast categories (M = 5.00), we found an interaction between relevant dimension and contrast, $F(1,132) = 10.911$, $p = .009$, $\eta^2 = .063$. Further analysis showed that this interaction was driven by a misperceived correlation between academic score and social score. When the relevant dimension was academic, and the target category was learned with a contrast category of lower values (i.e., the target category is being treated like an "honors dorm"), participants estimated social scores as being significantly lower ($M = 4.111$, $SD = 2.072$) than when the target category was in contrast to higher academic-scoring category (M = 6.920, SD = 1.741), $p < .001$. When the target category was thought of as more social, they estimated academic score as being significantly lower (M = 5.138, SD = 1.949), than when the category was thought of as less social ($M = 6.725$, $SD = 2.091$), $p = .049$.

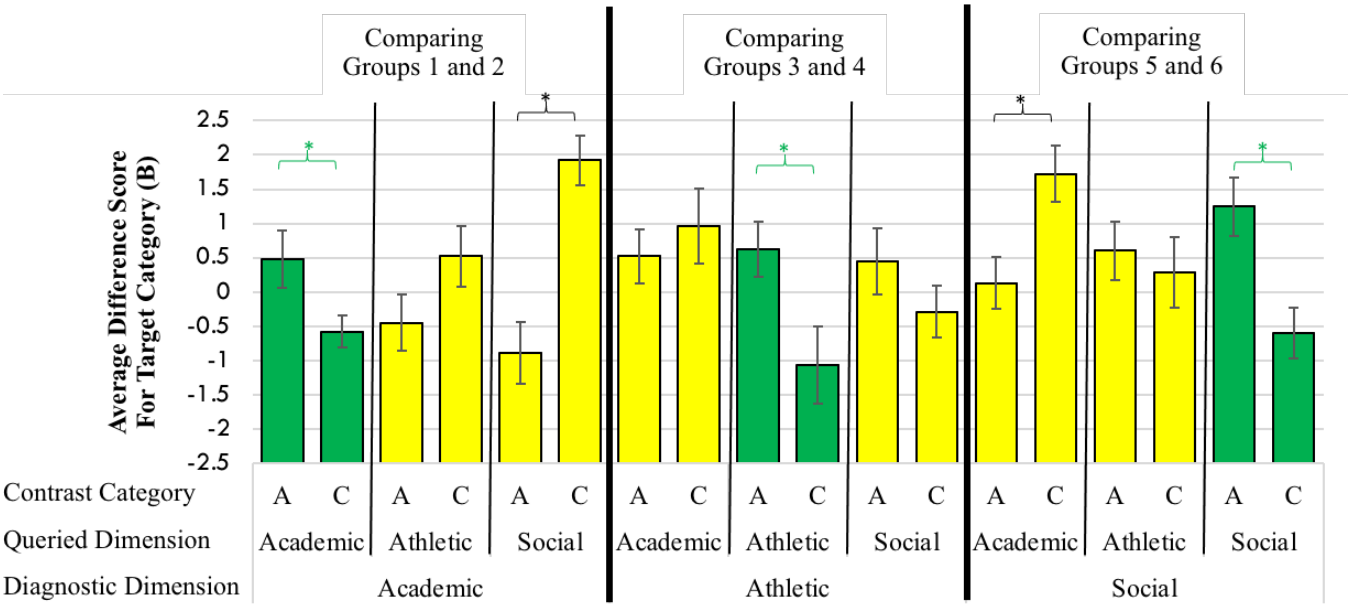

Figure 4: Average difference score for the target category (B) based on condition (contrast category and diagnostic dimension) and queried dimension. Bars/asterisks in green represent traditional contrast effects, as representation of queried dimensions were shifted in the direction away from contrast categories, which differed along that dimension. Black asterisks highlight significant contrast effects that have extended to non-diagnostic dimensions, reflecting a perceived correlation between academic and social dimensions.

Discussion

Participants were asked to learn the types of students who were assigned to different college dorms by comparing students in a dorm with central values along one dimension with students in a dorm with either higher or lower values along that dimension. As predicted, co-learned contrast categories with either higher or lower values along a relevant dimension had an influence on rate of learning and estimates of feature averages for a target category with central values. Further, in some cases these contrast effects actually extended to affect ratings along other dimensions. Specifically, students who were considered to have higher academic scores - based solely on the contrast category - were thought to be less social than those same students when they were compared to students with higher academic scores, and vice versa.

While the effect of contrast categories on dimensions of contrast has been well established in the literature, the reverberating effect of this contrast on *nondiagnostic* dimensions has not been well documented. In past research that carefully controls category representation based on classification relative to contrast categories (e.g., Davis & Love, 2010), representations of dimensions not relevant for classification have been largely unaffected. The current findings suggest that the context of learning about new social groups can in fact extend to affect features that are not useful for differentiating between categories if assumptions about relationships between features are pre-existing. For example, imagine you get a job at a new college and are learning about colleagues in the context of what academic department they are in. You may be more likely to conclude that professors in your Psychology Department are more creative if you share a building with business professors rather than art professors. The current research suggests that you may also develop the idea that psychology professors possess other traits you associate with creativity (e.g., open-mindedness, sensitivity), despite a lack of any evidence of those traits and despite those traits occurring equally across departments. Accordingly, these findings have implications for the perpetuation of stereotypes in the presence of neutral or disconfirming evidence.

At face value, these findings may seem obvious given long-standing research on the role of knowledge in concept learning and on basic biases in decision making (e.g., the representativeness heuristic). For example, we know that categories made up of coherent features that conform to established schema are learned faster than those that lack coherence (Murphy & Allopenna, 1994) or contain an uneven distribution of "crossover" features that are exceptions to categorical prototypes (Murphy & Kaplan, 2000). However, in this line of work interrelated features are typically baked into the learned category structure using discrete features that operate independent of a contrast category. For example, organizing a vehicle category around features like "made in Africa", "drives in jungles", and "is green" implies a theme even in the absence of distinguishing it from a category of "white" vehicles that "drive on glaciers". In the present work, the target category has average values along all dimensions and emerging themes or coherence only arises through the act of classifying examples relative to an alternative category. It is not immediately clear that these sort of contrast effects would activate themes or stereotypes that operate in the same way as if they were activated explicitly and independent of contrast.

We also know from the literature on decision making that probabilities can be misjudged in a variety of ways based on the perceived fit of an individual case to existing stereotypes (Tversky & Kahneman, 1972). A dorm full of students more athletic than their peers clearly activated stereotypical knowledge about what athletes are like and this knowledge either biased the formation of generalized knowledge about the probability of other traits or controlled their decision making in the moment they made a judgment about them. Again however, the fit of individuals to existing stereotypes was based on traits established solely *relative to a contrast category* and the question of how this sort of activation differs, if at all, is not well understood.

Questions remain about how far these contrast-based illusory correlations extend. For example, does the contrast specifically need to be made at the time of learning or would effects persist if one learned a new category in light of categories that have already been established? Further, in the current study illusory trait correlations were found to affect category representations based on training examples that did not possess any strong correlations along those traits. Further research should explore the extent to which illusory correlations bias representation of traits that are actually correlated in a way that is opposite of the assumptions behind illusory correlations. For example, how athletic does the more academic dorm need to be in order to counteract the assumed correlation between academic interest and lack of athletic pursuits?

It is worth noting that we did not ask directly about perceived correlations between features but indirectly inferred them from mean values abstracted from exposure to examples. We believe this methodology can offer advantages over explicit measures that may be more affected by conscious processing and social desirability biases. Using similar paradigms, future studies can investigate the effect of contrast on assumptions about other traits. For example, enhancing the stimuli by incorporating systematically varied faces would allow for exploration into assumptions about race, gender, attractiveness, and other physical traits. Generally, we view this approach as a complement to other attempts to occupy the somewhat rare middle ground between (1) classic studies of concept learning that focus on internal validity by rigorously controlling logical structure and task, and (2) studies in social cognition that may better address external validity by exploring more complicated social categories imbued with prior knowledge and ingrained biases.

References

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), 629–654. https://doi.org/10.1037/0278-7393.11.1-4.629

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition, 57*(3), 217–239. https://doi.org/10.1016/0010-0277(95)00669-X

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72*(3), 193–204. https://doi.org/10.1037/h0024670

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: Inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 216-226. https://doi.org/10.1037/0278-7393.30.1.216

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

Davis, T., & Poldrack, R. A. (2014). Quantifying the internal structure of categories using a neural typicality measure. *Cerebral cortex (New York, N.Y. : 1991)*, *24*(7), 1720–1737. https://doi.org/10.1093/cercor/bht014

Eiser, J. R. (1971). Categorization, cognitive consistency and the concept of dimensional salience. *European Journal of Social Psychology*, *1*(4), 435–454. https://doi.org/10.1002/ejsp.2420010404

Goldstone, R. (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. (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. (1996). Isolated and interrelated concepts. *Memory & Cognition, 24*(5), 608–628. https://doi.org/10.3758/BF03201087

Goldstone, R. L., Steyvers, M., & Rogosky, B. J. (2003). Conceptual interrelatedness and caricatures. *Memory & cognition*, *31*(2), 169–180. https://doi.org/10.3758/bf03194377

Hamilton, D. L., & Rose, T. L. (1980). Illusory correlation and the maintenance of stereotypic beliefs. *Journal of Personality and Social Psychology, 39*(5), 832–845. https://doi.org/10.1037/0022-3514.39.5.832

Harnad, S. (Ed.). (1987). *Categorical perception: The groundwork of cognition.* Cambridge University Press.

Hoffman, A. B. & Murphy, G. L. (2006). Category dimensionality and feature knowledge: When more features are learned as easily as fewer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 301-315. https://doi.org/10.1037/0278-7393.32.3.301

Kemler Nelson, D. G. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior, 23*, 734-759. https://doi.org/10.1016/s0022-5371(84)90442-0

Katz, P. A. (1963). Effects of labels on children's perception and discrimination learning. *Journal of Experimental Psychology, 66*(5), 423-428. https://doi.org/10.1037/h0045107

Krueger, J., & Rothbart, M. (1990). Contrast and accentuation effects in category learning. *Journal of Personality and Social Psychology, 59*(4), 651–663. https://doi.org/10.1037/0022-3514.59.4.651

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition, 43*(2), 266–282. https://doi.org/10.3758/s13421-014-0458-2

Levering, K., & Kurtz, K. J. (2010). Generalization in higher order cognition: Categorization and analogy as bridges to stored knowledge. In M. T. Banich & D. Caccamise (Eds.), *Generalization of knowledge: Multidisciplinary perspectives* (pp. 175–197). Psychology Press.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(3), 732–753. https://doi.org/10.1037/0278-7393.24.3.732

Minda, J. P. & Ross, B. H. (2004). Learning categories by making predictions: An investigation of indirect category learning. *Memory & Cognition, 32,* 1355-1368. https://doi.org/10.3758/bf03206326

Murphy, G. L. & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 904-909. https://doi.org/10.1037/0278-7393.20.4.904

Palmeri, T. J., & Nosofsky, R. M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 54A*(1), 197–235. https://doi.org/10.1080/02724980042000084

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgement. *British Journal of Psychology, 54*(2), 101–114. https://doi.org/10.1111/j.2044-8295.1963.tb00865.x

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131. https://doi.org/10.1017/cbo9780511809477.002