Title

Feature Ratings and Empirical Dimension-Specific Similarity Explain Distinct Aspect of Semantic SImilarity Judgments

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 40(0)

Authors

Iordan, Marius Catalin
Ellis, Cameron T
Lesnick, Michael
et al.

Publication Date

2018

Feature Ratings and Empirical Dimension-Specific Similarity Explain Distinct Aspects of Semantic Similarity Judgments

Marius Cătălin Iordan[1,2] (mci@princeton.edu), Cameron T. Ellis[3] (cameron.ellis@yale.edu)
Michael Lesnick[1], Daniel N. Osherson[2], Jonathan D. Cohen[1,2] ({mlesnick, osherson, jdc}@princeton.edu)
[1]Princeton Neuroscience Institute, [2]Department of Psychology, Princeton University, Princeton, NJ 08540 USA
[3]Department of Psychology, Yale University, New Haven, CT 06520 USA

Abstract

Predicting semantic similarity judgments is often modeled as a three-step process: collecting feature ratings along multiple dimensions (e.g., size, shape, color), computing similarities along each dimension, and combining the latter into an aggregate measure (Nosofsky, 1985). However, such models fail to account for over half of the variance in similarity judgments pertaining to complex, real-world objects (e.g., elephant and bear), even when taking into account their description along dozens of dimensions. To help explain this prediction gap, we propose a two-fold approach. First, we provide the first empirical evidence of a mismatch between similarity predicted by feature ratings and that reported by participants directly along individual dimensions. Second, we show that, surprisingly, separate sub-domains within directly reported dimension-specific similarities carry different amounts of information for predicting object-level similarity judgments. Accordingly, we show that differentially weighting directly reported dimension-specific similarity sub-domains significantly improves prediction of free (i.e., unconstrained) semantic similarity judgments.

Keywords: similarity judgments; semantics; representation; feature; dimension; object; category.

Introduction

Similarity judgments play a fundamental role in perception and reasoning, helping us to learn how new stimuli relate to previously learned categories, and to generalize this learning to novel situations. More specifically, similarity provides a metric for cognitive processes such as categorization, identification, and prediction (Ashby & Lee, 1991; Lambon Ralph, Jefferies, Paterson, & Rogers, 2017; Nosofsky, 1991; Rogers & McClelland, 2004; Tversky, 1977).

Similarity judgment has often been described as a mathematical function operating on descriptions of individual concepts along various dimensions; that is, in terms of their features, parts, and/or functions (Biederman, 1987; Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2016; Osherson, Stern, Wilkie, Stob, & Smith, 1991; Rogers & McClelland, 2004; Tversky & Hemenway, 1984). However, a consensus on the details of this function has remained elusive. More importantly, a major shortcoming of current theories of similarity is their inability to accurately predict the degree to which two complex real-world objects (e.g., two animals) are judged to be similar. For example, even if features along numerous dimensions (e.g., size, shape, color) are used to describe a collection of objects and compute the similarities among them, current models fail to capture more than half of the variance in directly reported similarity judgments (e.g., Osherson et al., 1991).

To address this problem, we focus on the role that dimension-specific similarity may play in contributing to similarity judgments regarding whole objects. We start with the observation that most existing models used to predict similarity judgments between objects ('How similar are these two objects?' e.g., Ashby & Lee, 1991; Osherson et al., 1991) from feature ratings of individual objects (e.g., 'How small/large is this object?') have used what appears to be a two-step procedure: a) collect empirical feature ratings for each object; b) represent each object as a vector of those feature ratings and use an element-wise operation to compute the distance between them as a prediction of their similarity. Whereas some models have differentially weighted separate dimensions when computing distance, in all cases the similarity of two objects within a given dimension has been assumed to be directly proportional to the overall distance between the two objects along that dimension. To the best of our knowledge, the validity of this assumption has not been empirically tested. That is, the similarity of complex real-world objects along a given dimension predicted by feature ratings has never been compared with direct (empirical) similarity judgments along that dimension (e.g., 'How similar are these two objects in terms of their *size*?'). We sought to address this gap by comparing estimates of dimension-specific similarities generated from feature ratings with empirically acquired dimension-specific similarity ratings. To the extent that these differed, we predicted that using the latter would improve prediction of object-level similarity ratings.

Most previous models have treated *all distances within a dimension equivalently* (e.g., two small objects are just as similar to each other as two big objects). However, the theory of structural alignment of cognitive representations (Gentner & Markman, 1994) provides evidence that this may not be the case. This is also suggested by work across multiple domains showing that psychological quantities often thought to be continuous or uniform are, in fact, better described by non-homogeneous, and even discrete scales (e.g., latent cause inference, Gershman & Niv, 2010; topic models, Blei, 2012; anchoring effects, Tversky & Kahneman, 1974; discrete representation of space in hippocampal place cell maps, Epstein, Patai, Julian, & Spiers, 2017). We explored whether this phenomenon may extend to the domain of feature representation and/or similarity judgment. One specific way in which behaviorally reported similarity along a particular dimension could differ from its computation based on feature ratings is the (potentially differential and/or non-monotonic) weight placed on specific values along that dimension. Specifically, we tested the hypothesis that similarities along individual dimensions may be best characterized by non-homogeneously weighted quanta, and that incorporating this insight into aggregate measures of similarity will improve predictions of object-level similarity judgements.

Materials and Methods

To test our hypotheses, we selected ten basic-level animals (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) and twelve feature dimensions and collected feature ratings for each animal on each dimension (Experiment 1), as well as unconstrained object-level similarity judgments (Experiment 2) and dimension-specific similarity judgments (Experiment 3) for each pair of animals.

Stimulus Set

We constructed a stimulus set comprising ten basic-level animals (bear, cat, deer, duck, parrot, seal, snake, tiger, turtle, whale) and twelve explanatory features (six objective features: size, domesticity, predacity, speed, furriness, aquatic-ness; six subjective features: dangerousness, edibility, intelligence, humanness, cuteness, interestingness). The objective features comprised a reasonable subset of features used throughout prior work on explaining similarity judgments (e.g., Osherson et al., 1991). However, little data has been collected how well subjective (and potentially more abstract or relational; Gentner, 1988; Medin, Goldstone, & Gentner, 1993) features can predict similarity judgments between pairs of real-world objects. Given that prior work has struggled to identify a subset of objective (and potentially more concrete) features that fully explain reported similarity judgments, we hypothesized that such subjective features may hold some of the missing variance and thus also help narrow the prediction gap illustrated in prior work.

For each of our ten animal categories, we selected nine 3-second videos showcasing the animal in its natural habitat. All videos were in color, contained the target animal as the largest and most prominent object in the scene, and were cropped to a size of 400x400 pixels from documentaries freely available online of minimum 720p quality (Fig. 1).

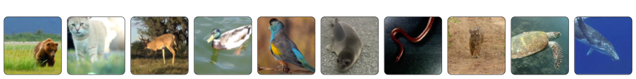

Figure 1. Examples of animal videos from the stimulus set.

Experiment 1: Feature Ratings

275 participants were recruited through Amazon Mechanical Turk in return for $0.50 payment. Participants were asked to rate each animal category (ten trials total, one animal shown per trial) on a randomly chosen dimension (e.g., 'How small/large is this animal?') on a discrete scale of 1 to 5 (1 = low feature value, e.g. 'small'; 5 = high feature value, e.g. 'large'). In each trial, they were shown three randomly selected videos from that category side by side and were given unlimited time to report a rating. Each participant saw each video at most once and the order of videos and categories was randomized across participants. Nineteen participants were excluded from the final analysis due to non-compliance with the instructions (e.g., RT below 200 ms for each trial, equal responses for all categories). We obtained average feature ratings for each animal by aggregating the ratings of the remaining participants (256: 19-26 per dimension).

Experiment 2: Object-Level Similarity

50 participants were recruited through Amazon Mechanical Turk in return for $2.00 payment. Participants were asked to report the similarity of each pair of animals ('How similar are these two animals?'; forty-five trials total) on a discrete scale of 1 to 5 (1 = not similar; 5 = very similar). In each trial, they were shown two randomly selected videos from different categories side by side and were given unlimited time to report a rating. Each participant saw each video at most once and the order of videos and categories was randomized across participants. Eight participants were excluded from the final analysis due to non-compliance with the instructions. We obtained an average object-level similarity for all animal pairs by aggregating the ratings of the remaining participants (42).

Experiment 3: Dimension-Specific Similarity

500 participants were recruited through Amazon Mechanical Turk in return for $2.00 payment. Participants were asked to report the similarity of each pair of animals (e.g., 'How similar are these two animals in terms of their size?') on a randomly chosen dimension (forty-five trials total) on a scale of 1 to 5 (1 = not similar; 5 = very similar). In each trial, they were shown two randomly selected videos from different categories side by side and were given unlimited time to report a rating. Each participant saw each video at most once and the order of videos and categories was randomized across participants. Thirty-three participants were excluded from the final analysis due to non-compliance with the instructions. We obtained average dimension-cued similarity measures for our animal pairs by aggregating the ratings of the remaining participants (467: 30-43 per dimension).

Results

From Feature Ratings to Object-Level Similarity

In Experiment 1, we collected a twelve-feature description for each animal in our stimulus set and we generated twelve rating-based dimension-specific similarity measures by computing the Euclidean distance between the ratings of each pair of animals on each dimension. Subsequently, we used a standard linear regression model where similarities

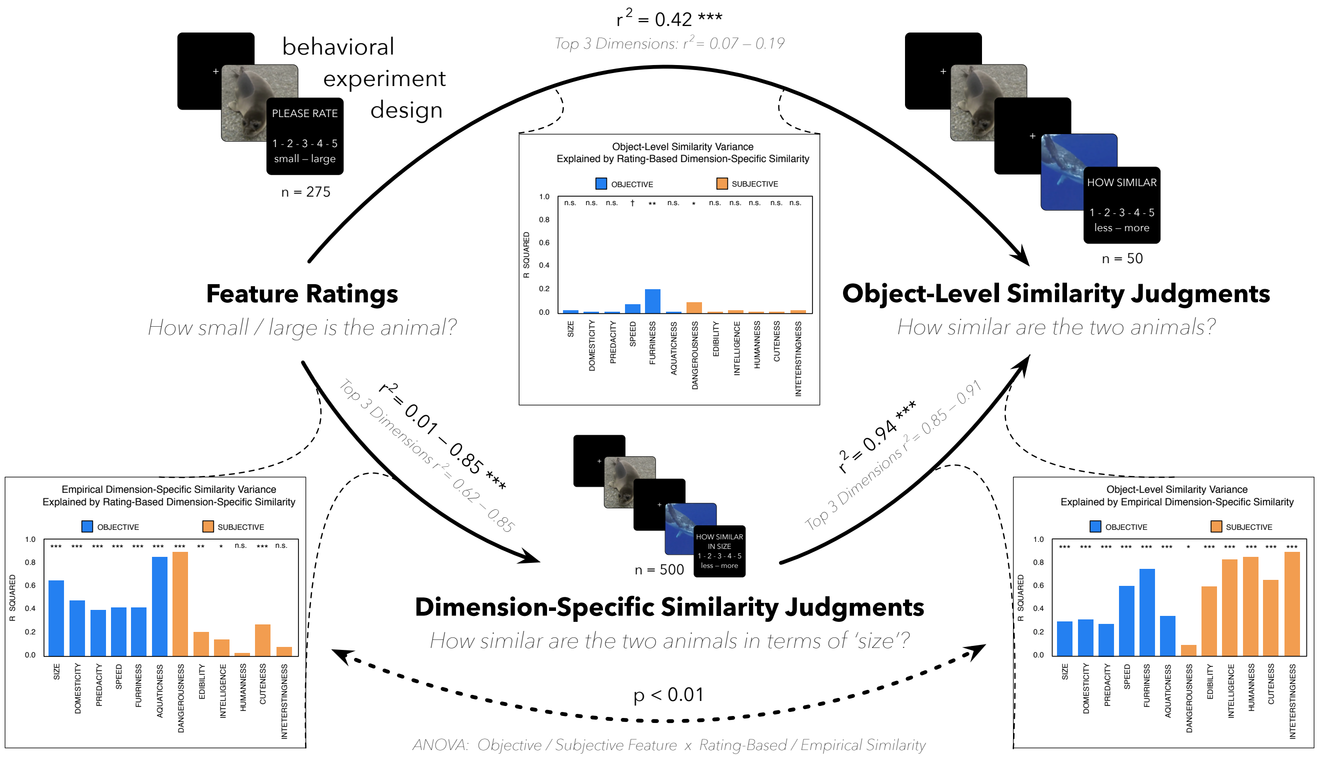

Figure 2. Experimental design and main variance explained results.

computed from feature ratings along each dimension were treated as separate predictors to test how well feature ratings were able to collectively explain object-level similarity ratings recorded in Experiment 2.

Consistent with prior work (Osherson et al., 1991), we found that by generating distances between animals from the feature ratings collected in Experiment 1 and subsequently optimally combining them into an aggregate measure (by weighting each dimension separately), we could predict object-level similarity collected in Experiment 2 reasonably well (r^2=0.42, p<0.01). By contrast, a similar procedure that combined each dimension equally had lower predictive power for object-level similarity (r^2=0.12, p<0.05) and each dimension individually did not usually predict object-level similarity above chance (Fig. 2, top; size: r^2=0.03, p=0.30; domesticity r^2<0.01, p=0.98; predacity: r^2<0.01, p=0.80; speed r^2=0.07, p=0.08; furriness: r^2=0.19, p<0.01; aquatic-ness r^2=0.09, p=0.04; dangerousness: r^2<0.01, p=0.77; edibility r^2<0.01, p=0.92; intelligence: r^2=0.04, p=0.20; humanness r^2<0.01, p=0.53; cuteness: r^2<0.01, p=0.61; interesting-ness r^2=0.04, p=0.19). Given the diversity of features tested, this suggests that similarity computed from feature ratings along individual dimensions cannot directly explain reported object-level similarity, unless combined into an aggregate measure. Even then, despite an over-representation of dimensions compared to objects being compared (twelve dimensions and ten animals), more than half of the variance in object-level similarity remains unexplained.

Empirical Dimension-Specific Similarity

To address this prediction gap, we hypothesized that a mismatch may exist between the similarity computed from feature ratings along various dimensions (e.g., using Euclidean distance) and the empirical dimension-specific similarity that participants would report if asked directly. To test this, in Experiment 3, we collected empirical dimension-specific similarity judgments for all pairs of animals (i.e., 'How similar are these two animals in terms of their size?') and used these judgments (instead of building a dimension-specific similarity measure from feature ratings along those dimensions) to predict object-level similarity.

We found that reported dimension-specific similarity was highly predictive of object-level similarity, not only at the aggregate level (r^2=0.94, p<0.01), but also significantly for most individual dimensions (Fig. 2, bottom right; size: r^2=0.30, p<0.01; domesticity r^2=0.32, p<0.01; predacity: r^2=0.28, p<0.01; speed r^2=0.61, p<0.01; furriness: r^2=0.76, p<0.01; aquatic-ness r^2=0.35, p<0.01; dangerousness: r^2=0.10, p=0.04; edibility r^2=0.60, p<0.01; intelligence: r^2=0.85, p<0.01; humanness r^2=0.86, p<0.01; cuteness: r^2=0.67, p<0.01; interesting-ness r^2=0.91, p<0.01). Furthermore, we observed a dichotomy between objective and subjective dimensions: the former contained less overlapping information about object-level similarity compared to the latter and, simultaneously, subjective dimensions were much more predictive of object-level similarity compared to objective dimensions (Fig. 2, bottom right; Fig. 3; ANOVA main effect of subjectivity, p<0.01).

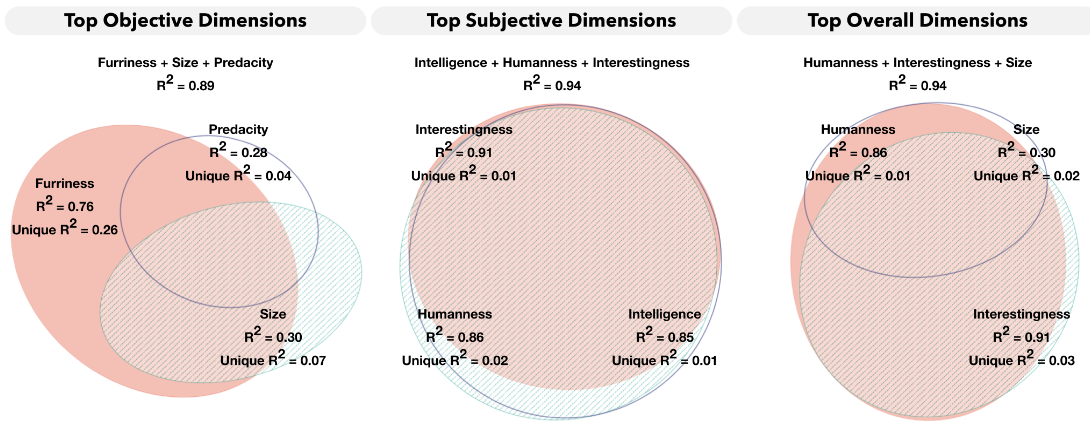

Figure 3. Variance explained by empirical dimension-specific similarity.

This suggests that the potential disconnect between similarity given by feature ratings along individual dimensions and reported object-level similarity may be due to limitations in building an accurate dimension-specific similarity measure from the feature ratings themselves. To test this possibility further, we measured how well empirical dimension-specific similarity could be predicted from the similarity generated from feature ratings along that dimension. There was high agreement between the two similarity measures (computed and empirical) for most dimensions considered in our experiments (Fig. 2, bottom left; size: r^2=0.62, p<0.01; domesticity r^2=0.51, p<0.01; predacity: r^2=0.38, p<0.01; speed r^2=0.40, p<0.01; furriness: r^2=0.40, p<0.01; aquatic-ness r^2=0.82, p<0.01; dangerousness: r^2=0.85, p<0.01; edibility r^2=0.18, p<0.01; intelligence: r^2=0.12, p=0.02; humanness r^2=0.01, p=0.49; cuteness: r^2=0.26, p<0.01; interesting-ness r^2=0.06, p=0.11). Furthermore, we found that feature ratings for objective dimensions were much more predictive of empirical dimension-specific similarity compared to subjective ones (ANOVA main effect of subjectivity, p<0.01), an effect directly opposite to the one between empirical dimension-specific similarity and object-level similarity (ANOVA interaction effect objective/subjective dimension x computed/empirical dimension-specific similarity, p<0.01).

Taken together, our results suggest that a significant portion of the missing explanatory power between similarity given by feature ratings and empirical object-level similarity judgments may lie in the intermediate step of constructing dimension-specific similarity. Moreover, this suggests the possibility that not all features are created equal in terms of how they relate both to the dimension-specific similarity they induce, and to how those intermediate dimension-specific similarity measures are subsequently combined to generate object-level similarity (as evidenced by the interaction effect we observed).

Non-Homogenous Within-Dimension Information

An alternative (or potentially complementary) explanation for the mismatch we observed between similarity derived from feature ratings and explicit similarity judgments between pairs of objects may arise from challenging the long held assumption that similarity information is uniformly distributed within each dimension. More specifically, consistent with the predictions of structural alignment theory (Gentner & Markman, 1994), it is possible that less similar (or conversely, highly similar) pairs of objects within a dimension may hold disproportionally more relevant information for computing overall similarity between those objects (for example, the fact that a mouse and a gerbil are almost identical in size may be much more informative for how similar they are judged, than the fact that a mouse and a rabbit have different sizes).

To test this hypothesis, we partitioned the similarity computed for each dimension based on feature ratings (Experiment 1) into a 'low similarity' half and 'high similarity' half and used each of these halves separately to predict empirical dimension-specific similarity judgments (Experiment 3). Across dimensions, we found that the two similarity halves behaved in an unsurprising manner: both low and high similarity were useful for predicting empirical dimension-specific similarity (Fig. 4, left; dimensions with significant prediction, Wilcoxon sign-rank test: predict empirical dimension-specific similarity, high vs. low, p=0.76).

We applied an analogous split-half analysis to the empirical dimension-specific similarity judgments (Experiment 3) to predict reported object-level similarity (Experiment 2). Here, we found strong evidence of within-dimension non-homogeneity: when predicting object-level similarity, the high similarity half of empirical dimension-specific similarity contained much more useful information than the lower similarity half (Fig. 4, right; dimensions with significant prediction Wilcoxon sign-rank test: predict all object-level similarity, high>low, p<0.01).

The observation that high and low empirical dimension-specific similarity contained differing amounts of information relevant to predicting object-level similarity suggests that it may be possible to improve the prediction of object-level similarity if the sub-domains of each dimension

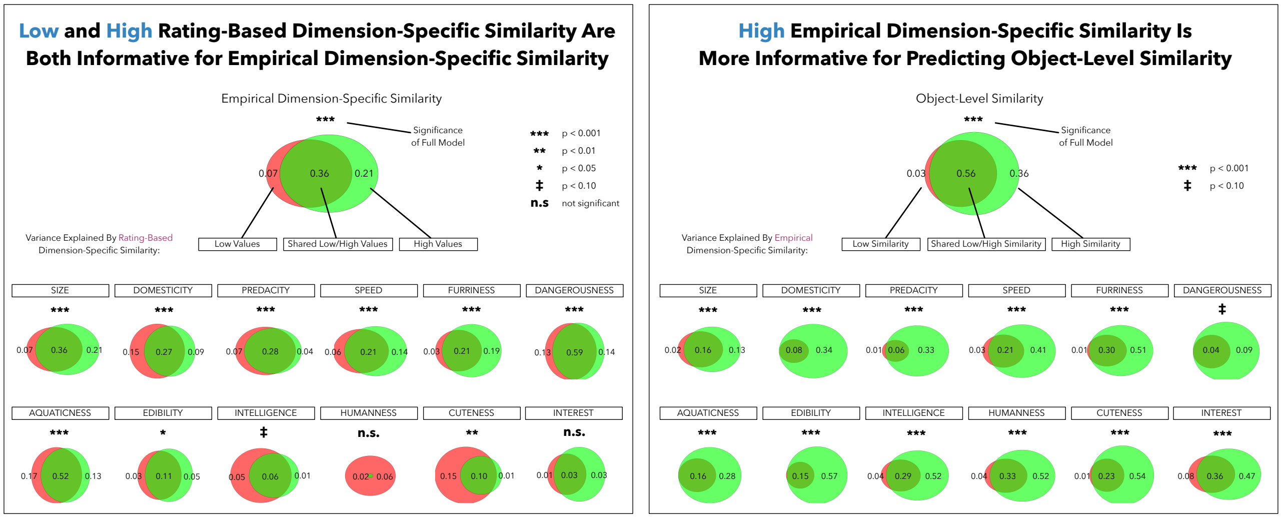

Figure 4. Non-homogenous information distribution within empirically reported dimension-specific similarity.

are weighted differently. Indeed, we found that using independent weights for each half of the empirical dimension-specific similarity explained more variance in object-level similarity (full < half dimensions, p=0.007; adjusted for the number of predictors). However, the same was not true for similarity computed from feature ratings along each dimension (full < half dimensions, p=0.250; adjusted for the number of predictors).

Taken together, these results suggest that separate sub-domains within each feature similarity continuum may contribute differently to how that particular feature explains similarity between the objects it describes. Furthermore, this provides further evidence for an important dichotomy between the two steps of the process of first building dimension-level similarity from feature ratings and subsequently combining this intermediate measure into a unified similarity judgment.

Discussion

The findings we report provide evidence that some of the missing explanatory power between feature ratings of individual objects and reported similarity between pairs of objects rests within the intermediate step of constructing an accurate dimension-specific similarity. Our design was the first, to our knowledge, to empirically measure the intermediate step of dimension-specific similarity for the purpose of quantifying its explanatory power for object-level semantic similarity judgments, compared to building models directly relating feature ratings to object-level similarity.

Furthermore, we showed that fine-grained distinctions between types of features (objective vs. subjective) interact across this intermediate computational step to diminish a direct predictive path from feature ratings to reported similarity. From prior observations (e.g., Gentner, 1988; Medin et al., 1993), we expect subjective, or potentially more relational features (e.g., humanness), to be more correlated with object-level similarity than objective, potentially more primitive features (e.g., size). In our work, however, we found that this relationship held only when using empirical dimension-specific similarity to predict object-level similarity, whereas the opposite was true when using similarity computed from feature ratings to predict its empirically observed counterpart. This dichotomy invites future work that investigates how similarity judgments differ across feature taxonomies in the context of empirical vs. computed dimension-specific similarities.

In most previous similarity models, usually a single weight was learned or posited for each object feature or dimension (e.g., Nosofsky, 1991; Osherson et al., 1991; Rogers & McClelland, 2004). However, we found that information within most features we examined was asymmetrically contained in distinct points along a putative continuum of representations (Fig. 4, right: high empirical dimension-specific similarity was an overwhelmingly better predictor of object-level similarity across a majority of dimensions, compared to low empirical dimension-specific similarity). This finding is consistent with the predictions of structural alignment theory applied to similarity between object pairs (Gentner & Markman, 1994) and, interestingly, this effect manifested most strongly when combining *empirical* dimension-specific similarities into an aggregate measure of object-level similarity, but less so when similarity was computed from feature ratings along those same dimensions. This suggests that participants may perform a systematic discounting of low similarity only after it has been already identified as such at the dimension level, and furthermore, that classical measures of dimension-specific similarity fail to take into account this effect. Alternatively, participants may be subject to an anchoring effect (Tversky & Kahneman, 1974) due to our experimental question emphasizing similarity over dissimilarity ('How similar are these two animals?'). While this account would still not fully explain the asymmetry of the effect across computed vs. empirical dimension-specific similarities, it may nevertheless be tested in future work by re-running the current experiments with the opposite prompt ('How *different* are these two animals?'). In any event, our findings suggest that – whether induced by local attentional effects, such as anchoring, or more stable representational factors – inhomogeneity in the influence of different points along a putatively continuous dimension may be an important factor in predicting object-level similarity.

Our results were derived under the assumption that Euclidean distance represents a reasonable measure for computing a dimension-specific similarity function from feature ratings. In a pilot version of our experiment, we additionally tested an exponential decay distance function (Shepard, 1988), a Gaussian similarity function (Nosofsky, 1985), and a city-block distance measure (Attneave, 1950; Garner, 1974). We chose Euclidean distance for our main experiment since all measures evinced qualitatively similar results, but Euclidean distance provided the highest average predictive power of all metrics we tested. Additionally, we did not observe any asymmetric similarity or comparison order effects (Medin et al., 1993; Nosofsky, 1991) pertaining to the sequential presentation of our stimuli during Experiments 2 and 3. A separate pilot experiment confirmed that the order of presentation for each animal within each trial/pair did not have any significant effect on the similarity ratings reported for that trial/pair).

One potential limitation of our study is the possibility that overall object-level similarity may exert a covert influence on dimension-specific similarity judgments collected in Experiment 3. Depending on the difficulty and unusualness of performing some dimension-level tasks (e.g., asking participants to actively consider 'humanness' of animals), participants may default to using object-level similarity as a prior and/or reporting a mixture of object-level and dimension-specific similarity as their overall judgment. Another potential limitation is a disparity between the Experiment 1 task (judgments involving single animals) and those of Experiments 2 and 3 (comparisons between two animals), which may affect the ability of similarity derived from the former to explain empirical similarity reported in the latter. We employed a design geared towards minimizing such effects (e.g., animals were presented sequentially in Experiments 2 and 3). However, the possibility remains that contextual effects and/or a meta-effect of actively performing a comparison versus a individual ratings may artificially increase the agreement between patterns of judgments in Experiments 2 and 3, compared to Experiment 1. Such complex effects should be further assessed in future experiments.

Given recent neuroimaging work suggesting an interaction between cognitive control (anterior cingulate cortex, ACC) and infero-temporal cortical regions in computing similarity judgments (Keung, Cohen, & Osherson, 2016; Lambon Ralph et al., 2017), our results provide an interesting hypothesis for elucidating the neural underpinnings of similarity judgments and their susceptibility to attention and other sources of bias. More specifically, the computations of dimension-specific similarities may be a precursor for computing object-level similarity, and thus the successful decomposition of the latter into a collection of the former may be measurable at the neural level as attention-induced perturbations in the representations of objects or semantic concepts (Çukur, Nishimoto, Huth, & Gallant, 2013). Furthermore, by showing evidence for discretization of information across multiple dimensions of similarity judgment, our work opens the possibility that semantic space may be internally represented as a cognitive map akin to ones theorized and investigated for spatial navigation in the hippocampus (Epstein et al., 2017). An interesting avenue for future work would be to test such a cognitive map model for computing similarity, potentially based on a semantic place cell analogy, where similarity judgments would operate as (potentially non-linear) transformations on distances between discrete points in dimension-specific feature maps.

Acknowledgements Funding: John Templeton Foundation.

References

Ashby, F. G., & Lee, W. W. (1991). Predicting similarity and categorization from identification. *Applied Psychological Measurement, 15,* 105-108.

Attneave, F. (1950). Dimension of similarity. *American Journal of Psychology, 63,* 516-556.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94(2),* 115-147.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55(4),* 77-84.

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience, 16,* 763-770.

Epstein, R. A., Patai, E. Z., Julian, J. B., & Spiers, H. J. (2017). The cognitive map in humans: Spatial navigation and beyond. *Nature Neuroscience, 20,* 1504-1513.

Garner, W. R. (1974). *The processing of information and structure*. New York, NY: Wiley.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development, 59,* 47-59.

Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science, 5(3),* 152-158.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology, 20(2),* 251-256.

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General, 145(1),* 82-94.

Keung, W., Cohen, J. D., & Osherson, D. N. (2016). Influence of cognitive control on semantic representation. *bioRXiv,* doi.org/10.1101/067553.

Lambon Ralph, M. A., Jefferies, E., Patterson, K, & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience, 18,* 42-55.

Medin, D., Goldstone, R., & Gentner, D. (1993). Respects for similarity. *Perception & Psychophysics, 38,*415-432.

Nosofsky, R. M. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics, 38,*415-432.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology, 23(1),* 94-140.

Osherson, D. N., Stern, J., Wilkie, O., Stob, M., & Smith, E. E. (1991). Default probability. *Cognitive Science, 15,* 251-269.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing.* Cambridge, MA: MIT Press.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8(3),* 382-439.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317-1323.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84(4),* 327-352.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185(4157),* 1124-1131.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General, 113(2),* 169-193.