AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 3, 573–589
https://doi.org/10.1037/xge0001533

The Universal Law of Generalization Holds for Naturalistic Stimuli

Raja Marjieh[1], Nori Jacoby[2], Joshua C. Peterson[3], and Thomas L. Griffiths[1, 3]

[1] Department of Psychology, Princeton University

[2] Computational Auditory Perception Group, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany

[3] Department of Computer Science, Princeton University

Shepard's universal law of generalization is a remarkable hypothesis about how intelligent organisms should perceive similarity. In its broadest form, the universal law states that the level of perceived similarity between a pair of stimuli should decay as a concave function of their distance when embedded in an appropriate psychological space. While extensively studied, evidence in support of the universal law has relied on low-dimensional stimuli and small stimulus sets that are very different from their real-world counterparts. This is largely because pairwise comparisons—as required for similarity judgments—scale quadratically in the number of stimuli. We provide strong evidence for the universal law in a naturalistic high-dimensional regime by analyzing an existing data set of 214,200 human similarity judgments and a newly collected data set of 390,819 human generalization judgments ($N = 2{,}406$ U.S. participants) across three sets of natural images.

Public Significance Statement

Humans constantly form generalizations, whether when trying to identify the color of an object or reasoning about which action to take based on past experiences. Understanding how generalizations relate to underlying psychological representations is a core problem in cognitive science. The universal law of generalization is a fundamental hypothesis concerning the nature of this relationship which states that the strength of generalization between two stimuli should decay as a universal exponential function of their psychological distance. While extensively studied, evidence for the universal law comes from small data sets and artificial stimuli that are very different from the real world. Our work is the first to provide strong evidence for the universal law in a high-dimensional naturalistic domain by collecting and analyzing 605,019 human similarity and generalization judgments for natural images.

Keywords: generalization, similarity, perception, natural images, representations

Every day, humans interact with complex perceptual objects that vary in modality, structure, and function. Whether deciding when to cross a street, recognizing the face of a friend, or trying to determine whether a novel fruit will taste good, we need to form meaningful generalizations from past perceptual experiences. This problem of generalization is arguably one that we share with all intelligent species, something that led Roger Shepard to propose a candidate for the first universal law of psychology (Shepard, 1987). Shepard's universal law of generalization—intended to hold for intelligent entities anywhere in the universe—asserts that the extent to which a property is generalized from one stimulus to another should decrease as a concave function (usually exponential) of the distance between those stimuli in psychological space. This idea has been elaborated upon in Bayesian models of cognition (Tenenbaum & Griffiths, 2001), and linked to information-theoretic principles such as maximum entropy (Myung & Shepard, 1996), Kolmogorov complexity (Chater & Vitányi, 2003), and efficient coding (Sims, 2018).

Implicit in Shepard's proposal is the idea that it is possible to represent perceptual stimuli in a psychological space—typically a low-dimensional representation where the similarity between two stimuli decreases with their distance. While this idea is controversial (e.g., Peer et al., 2021; Tversky, 1977; Tversky & Hutchinson, 1986), Shepard showed that such spaces can capture the similarity relationships between a variety of simple perceptual stimuli. He proposed a procedure, known as multidimensional scaling (MDS), for uncovering the structure of mental representations from behavioral data (Shepard, 1962, 1980; Steyvers, 2006). Given a set of stimuli, the

Raja Marjieh https://orcid.org/0000-0001-8156-1333

This work was supported by Grant 61454 from the John Templeton Foundation. A preliminary version of these results was presented at the 2023 Annual Meeting of the Society of Experimental Psychologists. The authors thank Laurence Maloney for suggesting the analysis based on Tversky and Hutchinson (1986). The authors declare no competing interests. All data and code are available at the following link: https://osf.io/rbkgh.

Raja Marjieh contributed to software, visualization, and writing–original draft. Thomas L. Griffiths contributed to funding acquisition. Raja Marjieh, Nori Jacoby, and Thomas L. Griffiths contributed equally to conceptualization and methodology. Raja Marjieh and Joshua C. Peterson contributed equally to investigation. Raja Marjieh, Nori Jacoby, Joshua C. Peterson, and Thomas L. Griffiths contributed equally to writing–review and editing. Raja Marjieh and Nori Jacoby contributed equally to formal analysis.

Correspondence concerning this article should be addressed to Raja Marjieh, Department of Psychology, Peretsman-Scully Hall, Princeton University, Princeton, NJ 08540, United States. Email: raja.marjieh@princeton.edu

procedure begins by constructing a similarity matrix between all stimulus pairs, for example, by collecting similarity judgments or confusion probabilities, and then applying an iterative algorithm that embeds those stimuli in a psychological space (typically Euclidean) such that similar stimuli are mapped to nearby points.

Having mapped stimuli to points in a psychological space, it becomes possible to test the universal law. While Shepard's (nonmetric) MDS method assumes similarity decreases with distance, it does not specify the form of that function. This follows from the fact that, unlike metric MDS which seeks an exact mapping between similarity s and distance d (i.e., $1 - s = d$), nonmetric MDS only imposes the ordinal relations induced by s which, in turn, allows it to find more flexible linking functions $s = f(d)$ between similarity and distance, when d is required to arise from distances in a Euclidean space (since distances are highly constrained by ordinal relations up to a scaling factor when the number of points is large; Shepard, 1966). Indeed, Shepard (1962) demonstrated that this flexibility of nonmetric MDS allows it to recover arbitrary monotonic linking functions between distance and similarity. By further analyzing the abstract question of how an ideal organism should decide whether two stimuli shared a given property, Shepard (1987) showed that this function should be concave. Mathematical analysis of a variety of different assumptions about the distribution of properties in psychological space showed that generalization typically decreased as an exponential function of distance. Shepard then demonstrated that this theoretical relationship held for a wide array of stimuli that had been embedded into a psychological space via MDS, including geometric shapes, phonemes, colors (in both humans and pigeons), and even Morse code signals.

Despite the success of Shepard's account, two clear limitations remain. First, for a set of N stimuli, MDS requires on the order of N^2 pairwise comparisons to construct a full similarity matrix which, as the number of stimuli increases, necessitates a large amount of human data. For example, a set of 100 stimuli would require on the order of 10,000 similarity judgments, without even including any repetitions to ensure data quality. This bottleneck has recently propelled a line of research aimed at finding cheaper approximations for human similarity matrices (Jha et al., 2023; Marjieh et al., 2023; Roads & Love, 2021). Second, and in part as a result of the first limitation, most of the evidence for the universal law comes from studies that are limited to low-dimensional artificial stimuli and small stimulus sets (Cheng, 2000; Ghirlanda & Enquist, 2003; Shepard, 1987; Sims, 2018). Even though more recent work such as that of Sims (2018) has considered somewhat richer stimuli such as synthesized instrument timbres and vibrotactile patterns, these were still limited to small data sets on the scale of 10–20 stimuli. These limitations make it hard to draw conclusions about the status of the universal law of generalization in the high-dimensional regime of real-world stimuli, especially as fundamental problems in psychology continue to be reshaped by large-scale behavioral studies (see, e.g., Awad et al., 2018; Battleday et al., 2020; Marjieh et al., 2022; Peterson et al., 2021).

To address this gap, we leveraged recent advances in online recruitment as well as the availability of naturalistic image data sets to test the universal law of generalization in a high-dimensional setting. Specifically, we considered a data set of similarity judgments over three sets of images recently collected by Peterson et al. (2018) where each data set comprised 120 images from a given natural category, namely, animals, fruits, and vegetables. This data set consisted of 214,200 human judgments. To account for the different ways in which similarity scores can be constructed, we augmented this data set with a newly collected set of generalization judgments where participants rated how likely it is a certain blank property (Kemp & Tenenbaum, 2009; Osherson et al., 1990; e.g., having an enzyme) generalizes from one stimulus to another. The latter data set comprises 390,819 generalization judgments from 2,406 online participants. We used these data to directly test the universal law of generalization in this high-dimensional large-scale regime.

Method

Our approach builds on advances in large-scale online recruitment and experiment design to exhaustively estimate similarity matrices over naturalistic stimuli by directly scaling up pairwise judgment elicitation. For stimuli, we focused on natural images (Figure 1A) for three reasons, namely, (a) they strike a balance between being perceptually complex and being intuitive and widespread across cultures, (b) they can be easily embedded within an online study, which facilitates crowdsourcing, and (c) high-quality sets of natural images along with accompanying behavioral data are available in the literature (Jha et al., 2023; Peterson & Griffiths, 2017; Peterson et al., 2018). As for the paradigm, we used simple pairwise judgment elicitation on a Likert scale with two complementary types of human judgments that are common to the study of representations, namely, direct similarity judgments that answer the query "How similar are the animals in the following two images?" (Figure 1B; Peterson et al., 2018; Shepard, 1980), and generalization judgments that answer the query "If the animal in Image 1 has enzyme X321, how likely is it that the animal in Image 2 has it too?" (Figure 1C; Kemp & Tenenbaum, 2009; Osherson et al., 1990). In the latter case, we used a fictitious enzyme name so as to prevent participants from resorting to any technical knowledge. Recent work on similarity judgments for images has also used an alternative paradigm where participants judge which of three images is the odd one out (Hebart et al., 2020). However, we chose to use pairwise similarity judgments to maximize the correspondence with the paradigm used by Shepard, and because triplets scale cubically in the number of stimuli making them even harder to scale (without further assumptions about deriving pairwise similarity from triplets).

Stimuli

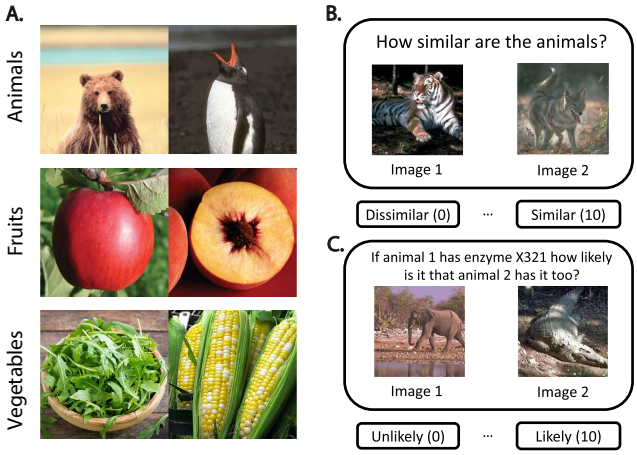
We used sets of images from three natural categories, following Peterson et al. (2018). Each set comprised 120 images from one of the following categories: animals, fruits, and vegetables (see examples in Figure 1A). In addition, these data sets were supplemented with full 120×120 symmetric human similarity matrices s_{ij} where each entry corresponds to an aggregate similarity score between an image i and an image j in the range 0–1, where a value of 0 indicates complete dissimilarity, and a value of 1 indicates complete similarity. Each such similarity matrix was constructed using 71,400 human judgments from a pool of approximately 1,200 U.S. participants recruited on Amazon Mechanical Turk (AMT; Peterson et al., 2018).

Participants

Participants for the generalization tasks were recruited online via AMT subject to the following criteria to ensure data quality: (a) participants must be at least 18 years of age, (b) they must reside in the

Figure 1

Example Stimuli and Schematics of the Different Behavioral Paradigms



Note. (A) Example images from three natural categories, namely, animals, fruits, and vegetables, and schematics of the two elicitation queries used in the present work, namely, (B) direct similarity judgments, and (C) generalization judgments. Animal examples are adapted from Berkeley Segmentation Dataset and Benchmark (BSD), "A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," by D. Martin, C. Fowlkes, D. Tal, and J. Malik, *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001* (Vol. 2, pp. 416–423), 2001, Copyright 2001 by IEEE. Fruit samples are adapted from *Red Apple With Leaf*, by Paul Neo, 2013 (https://commons.wikimedia.org/wiki/File:Red_apple_with_leaf.jpg). In the Public Domain (CC0); *Autumn Red Peach*, by Jack Dykinga, U.S. Department of Agriculture, 2013 (https://en.m.wikipedia.org/wiki/File:Autumn_red_peach.jpg). In the Public Domain (CC0). Vegetable samples are adapted from *Two-Color Corn*, by Rosana Prada, 2007 (https://www.flickr.com/photos/zanastardust/1303616796). CC BY 2.0; *Fresh Arugula Salad*, by Pilipphoto on Adobe Stock (https://stock.adobe.com/images/fresh-arugula-salad/67887886). Copyright by Adobe Stock Extended License (purchased by authors). See the online article for the color version of this figure.

United States, and (c) they must have an approval rate of 95% or higher on AMT. The recruitment process was performed using the Dallinger[1] platform for experiment hosting, and the experimental session was programmed using PsyNet, a framework for online experiment design that is built on top of Dallinger (Harrison et al., 2020). Overall, $N = 2{,}406$ participants completed the studies, and they were paid \$12/hour for their participation. Specifically, the $N = 773$ participants in the animals condition had an age range of 21–78 years ($M = 38.3$, $SD = 11.0$), the $N = 833$ participants in the fruits condition had an age range of 19–70 years ($M = 39.3$, $SD = 11.0$), and the $N = 800$ participants in the vegetables condition had an age range of 20–77 years ($M = 39.1$, $SD = 10.9$). The sample size was selected such that each image pair received an average of nine ratings to match that of the data sets of Peterson et al. (2018).

Procedure

After completing a consent form, participants received the following instructions "In this experiment, we are studying how people form generalizations. In each trial of this experiment, you will be presented with two images of animals/fruits/vegetables. One of the animals/fruits/vegetables will possess a certain property, and your task will be to judge based on that information how likely it is that the second animal/fruit/vegetable has that property. You will have 11 response options, ranging from 0 (*not likely at all*) to 10 (*very likely*). Choose the one you think is most appropriate." Participants then proceeded to the main experiment where they were presented with image pairs followed by the prompt "If the animal/fruit/vegetable in the left image has enzyme X132, how likely is it that the animal/fruit/vegetable in the right image has it too?" (see schematics in Figure 1C). Overall, 390,819 judgments were elicited with each participant providing up to 200 judgments. The procedure in the similarity paradigm of Peterson et al. (2018) was analogous. Participants rated the similarity between pairs of images on a Likert scale ranging from 0 (*not similar at all*) to 10 (*very similar*; Figure 1B; see Peterson et al., 2018 for additional details).

Data Analysis

From Generalization to Similarity

To convert generalization scores into similarity matrices the following preprocessing was applied. First, the responses of individual participants were z-scored (within participants) to account for different usage of the response scale across participants. Then, the z-scored ratings were averaged across participants to produce a single score per stimulus pair. The summarized z-scores were then converted into generalization probabilities p_{ij} by passing them through a cumulative normal distribution. Finally, to derive symmetric similarity matrices s_{ij} we applied Shepard's similarity formula $s_{ij} = \sqrt{p_{ij}p_{ji}/p_{ii}p_{jj}}$ (Shepard, 1987).[2] In practice, we noticed that a few of the diagonal probabilities p_{ii} were smaller than their off-diagonal counterparts which resulted in a generalization score that is >1 and hence a negative entry in the distance (dissimilarity) matrix ($\Delta_{ij} = 1 - s_{ij}$), likely due to noise in the similarity estimates. Since these entries constitute only extremely small fraction of the data (0.8%), we truncated the diagonal values by setting p_{ii} to one, similar to Peterson et al. (2018).

MDS

Given a dissimilarity matrix Δ_{ij}, MDS embeddings z were obtained using the manifold.MDS method from the scikit-learn Python library (Pedregosa et al., 2011) with a maximum iteration limit of 10,000 and a convergence tolerance of 10^{-100}. Embeddings were computed in two steps: first metric MDS was applied to get an initial embedding which was then used to initialize nonmetric MDS. We chose a $d = 4$ dimensionality for the embedding space based on an MDS stress curve analysis (shown in Figure A1; for visualization purposes only we used $d = 2$ in the figures below) whereby the first dimension d for which all stress values across all data sets dropped below 0.2 was selected (a standard threshold above which MDS fit is deemed poor; Kruskal, 1964). Finally, to construct generalization gradients we computed Euclidean distance

[1] https://dallinger.readthedocs.io/en/latest/.

[2] Note that this formula does not change the similarity matrices of Peterson et al. (2018) since $\sqrt{s_{ij}s_{ji}/s_{ii}s_{jj}} = \sqrt{s_{ij}^2} = s_{ij}$ due to the fact that $s_{ij} = s_{ji} \geq 0$ and that $s_{ii} = s_{jj} = 1$.

between all MDS embedding vectors $d_{ij} = \|z_i - z_j\|_2$ and combined them with their corresponding similarity scores s_{ij} to produce the two-dimensional set $D = \{(d_{ij}, s_{ij})\}$. We analyzed the resulting generalization gradients in two complementary ways, namely, by directly fitting curve models to the raw set D, and by fitting them to a binned version of D. The former served as a conservative test, and the latter as more balanced one meant to evaluate the average curve and to take into account the fact that different regions of the generalization gradient have different densities (e.g., high similarity pairs are much less common than low similarity pairs which can overemphasize the tail of the gradient). The binning was done by dividing the distances $\{d_{ij}\}$ into 100 bins and then computing the average d_{ij} and s_{ij} within each bin and their standard errors.

Model Fitting and Evaluation

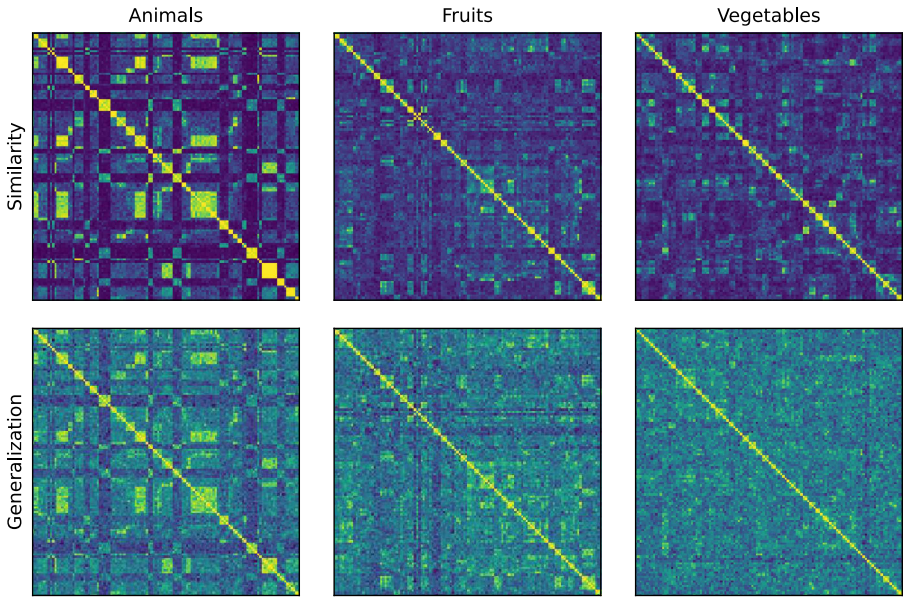
To test the universal law, we evaluated the extent to which an exponential function of the form $g(x) = ae^{-bx} + c$ could account for the generalization gradients D relative to four other models of increasing complexity, namely, a simple linear model $g(x) = ax + b$, a quadratic model $g(x) = ax^2 + bx + c$ (same complexity as exponential but with the option of being either concave of convex), a Gaussian model $g(x) = ae^{-bx^2} + c$ (a well-known alternative often discussed in the literature; Chater & Vitányi, 2003), and a flexible generalized additive model (GAM), that is, a model of the form $g(x) = \alpha + \sum_i \beta_i f_i(x)$ where $f_i(x)$ is a basis of cubic splines (Hastie et al., 2009), as well as the intrinsic interrater variability of the data. To fit the exponential, linear, quadratic, and Gaussian models we used the curve_fit least squares optimizer in scipy, and to fit the GAM we used the LinearGAM method of the pygam package (Servén & Brummitt, 2018) which by default uses 20 basis functions and optimizes the model parameters in a cross-validated grid-search manner. For model evaluation, and to accommodate both for the possibility of overfitting and to adjust for degrees of freedom, we performed a split-half bootstrap analysis whereby 100 data splits were produced by randomly dividing the ratings per image pair in half and then producing two generalization gradients D_{h_1} and D_{h_2} to which the model was fitted yielding two sets of predictions $\{s'_{ij}\}_{h_1}$ and $\{s'_{ij}\}_{h_2}$. We then computed the following Pearson correlation coefficients between the data–model sets $\{s_{ij}\}_{h_1}$, $\{s'_{ij}\}_{h_1}$ and $\{s_{ij}\}_{h_2}$, $\{s'_{ij}\}_{h_2}$: r_{dd} data–data correlation, r_{mm} model–model correlation, r_{dm} data–model correlation (there are two splits for each randomized split half, and thus there are two ways to compute this which we averaged), and $r_{\text{c}} = r_{\text{dm}}/\sqrt{r_{\text{dd}} r_{\text{mm}}}$ the data–model correlation corrected for attenuation (Jensen, 1998). In addition, we provide the coefficient of determination (variance explained [VE]) R^2 for each of the fitted models. Finally, since we are specifically interested in how models perform relative to the exponential model, we bootstrapped the difference in the coefficient of determination $\Delta R^2 = R^2_{\text{exponential}} - R^2_{\text{model}}$ on each training half ($\Delta R^2_{\text{train}}$) as well as its corresponding test half (ΔR^2_{test}). This way we can assess both the relative quality of fit as well as the model's generalization ability (i.e., penalize complex models that are too flexible and simple models that are too rigid). We note that in computing all metrics we excluded trivial self-similarity points ($d = 0$, $s = 1$) to prevent artificial inflation of values.

Transparency and Openness

All data and analysis code considered in the present work, as well as all necessary code for reproducing the online behavioral experiments,

Figure 2

Similarity Matrices Over the Different Domains of Natural Images and the Two Judgment Elicitation Tasks Considered, Namely, Direct Similarity Judgments and Generalization Judgments



Note. See the online article for the color version of this figure.

are made publicly available in the following Open Science Framework repository: https://osf.io/rbkgh (see Method for additional details regarding data analysis). This work was not preregistered. All participants provided informed consent prior to participation in accordance with an approved Princeton University Institutional Review Board (Protocol 10859).

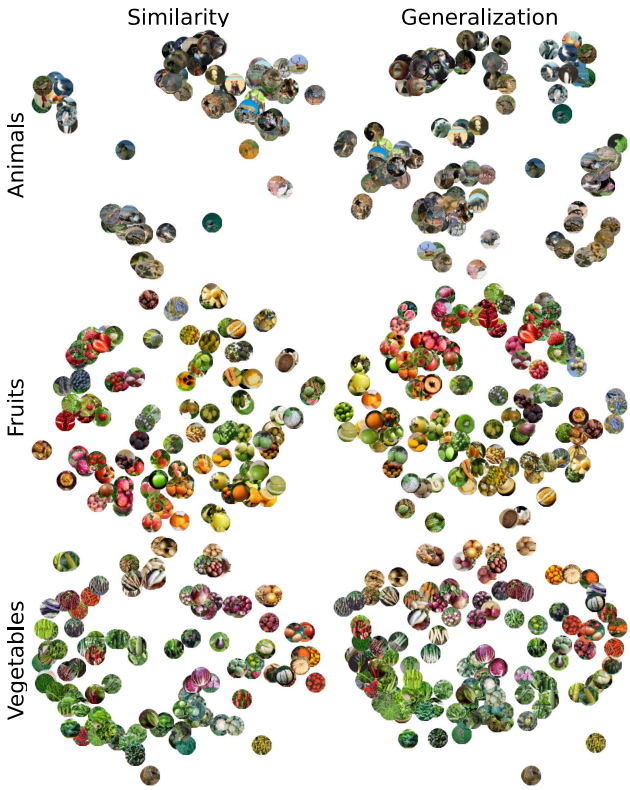
Results

The average similarity matrices over the different domains and tasks are summarized in Figure 2. The first thing to observe is that the generalization and similarity judgment tasks yield results that are significantly correlated across domains, with a Pearson correlation of $r = .71$ (95% confidence interval [CI] [.69, .74]) for animals, $r = .55$ (95% CI [.52, .58]) for fruits, and $r = .36$ (95% CI [.33, .40]) for vegetables (CIs bootstrapped over participants with 100 repetitions). This is consistent with the expectation that generalization over blank properties and similarity judgments capture shared variance (Kemp & Tenenbaum, 2009).

Next, to get a better sense of the psychological content of those spaces, we visualized their two-dimensional MDS solutions in Figure 3. All three domains revealed a semantically structured organization of the stimuli (a large high-resolution version of the figure is provided in the OSF repository). In the case of animal images, distinct and interpretable organization schemes emerged, corresponding to animal categories such as herbivores, carnivores, amphibians, reptiles, and birds. In the case of fruits and vegetables, the distribution is more continuous, with color serving as a clear semantic axis, with interpretable subclasses occupying different areas of the space such as citrus fruits and berries in the case of fruits, and whether the vegetable grows above or below the ground in the case of vegetables. These results are consistent with the findings of Peterson et al. (2018) in the case of similarity and extend them to generalization, implying again that both tasks are capturing shared semantic content.

Figure 3

Two-Dimensional MDS Embeddings for the Similarity and Generalization Data With the Raw Image Stimuli From the Different Natural Categories Overlaid



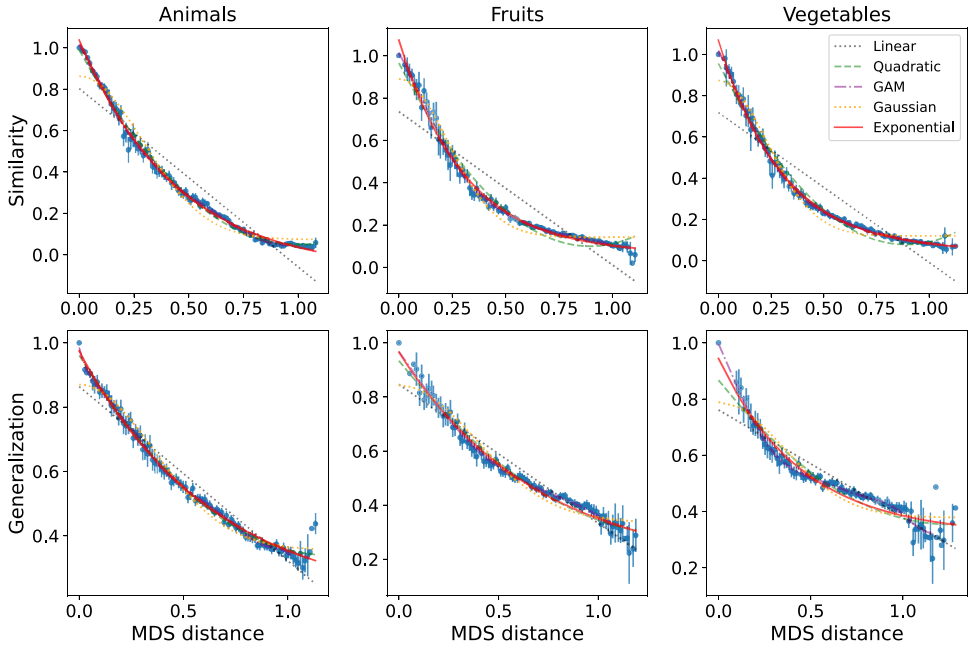
Note. MDS = multidimensional scaling. See the online article for the color version of this figure.

We are now ready to analyze the generalization gradients for each of the conditions and test to what extent they can be explained by the different models. The average binned gradients along with their optimal fit for all models are shown in Figure 4 (see Method; raw gradients are shown in Figure A2; explicit fitted parameter values and their CIs are provided in Tables A1–A4 and B1–B4 for the raw and binned analyses, respectively). As can be seen, the scatter points appear to follow a concave trend, with the similarity data in particular tightly tracking the exponential curve, which also overlaps substantially with the quadratic, Gaussian (except at short distances), and GAM models. The linear model, on the other hand, appears to find some intermediate compromise due to its limited flexibility. To quantify this, we provide the full list of evaluation metrics on the raw gradients in Tables 1 and 2 (see Tables A5 and A6 for additional metrics; see also Tables B5–B8 for the binned equivalent). The exponential function provides an excellent model for the data with an average model–data Pearson correlation of $r_c = .96$ (corrected for attenuation, see Method). To see where the exponential model stands with respect to the different models in each condition, we performed a cross-validation analysis whereby we bootstrapped the difference in VE relative to the exponential model $\Delta R^2 = R^2_{\text{exponential}} - R^2_{\text{model}}$ across training ($\Delta R^2_{\text{train}}$) and test ($\Delta R^2_{\text{test}}$) split-halves (see Method). Starting from the domain of similarity, we found that the exponential model outperformed the linear model in all three domains, with $\Delta R^2_{\text{train}}$ 95% CIs given by [.06, .08], [.12, .15], [.12, .15] for animals, fruits, and vegetables, respectively (positive values favor exponential in our definition). The same holds for the corresponding ΔR^2_{test} CIs (i.e., when penalizing for complexity), [.07, .08], [.13, .15], [.13, .15]. As for the quadratic solution, the models performed practically the same (with a slight boost for the exponential) with $\Delta R^2_{\text{train}}$ CIs given by [−.002, .001], [.006, .020], and [.008, .016] for animals, fruits, and vegetables, respectively (and likewise for the corresponding ΔR^2_{test} [−.001, .001], [.013, .023], and [.013, .019]). Crucially, however, all quadratic solutions converged on concave curvature with strictly positive second derivatives $g''(x) = 2a > 0$ with CIs [1.77, 1.85], [2.36, 2.24], and [2.22, 2.31] (see Table A3) consistent with the universal law hypothesis. Next, for the Gaussian solution, we have $\Delta R^2_{\text{train}}$ CIs [.01, .02], [.01, .03], and [.01, .03] and ΔR^2_{test} CIs [.01, .02], [.02, .03], and [.02, .03] for animals, fruits, and vegetables, which suggest a small but robust preference for the exponential model. This small difference is expected as the Gaussian solution is largely concave except for the near-zero region. Finally, for the flexible GAM model, we found that it was unable to meaningfully improve on the exponential model despite its flexibility, with $\Delta R^2_{\text{train}}$ CIs given by [−.008, .000], [−.010, −.002], [−.008, −.001], and ΔR^2_{test} CIs given by [−.005, −.001], [−.004, .002], [−.004, .001], for animals, fruits, and vegetables, respectively.

As for the generalization data, we observed a similar pattern, namely, the exponential model outperformed the linear ($\Delta R^2_{\text{train}}$ CIs, [.016, .029], [.009, .028], [.006, .030], and ΔR^2_{test} CIs, [.021,

Figure 4

Generalization Gradients Across Domains of Natural Images and Tasks With the Optimal Model Fits Overlaid



Note. Error bars indicate 95% confidence intervals. GAM = generalized additive model; MDS = multidimensional scaling. See the online article for the color version of this figure.

.032], [.021, .034], [.030, .054], for animals, fruits, and vegetables, respectively). Likewise, the quadratic model performed on par with the exponential model ($\Delta R^2_{\text{train}}$ CIs, [.000, .002], [.002, .005], [.005, .014], and ΔR^2_{test} CIs, [.000, .002], [.005, .008], [.021, .028]), and was strictly concave on all domains ($g''(x) = 2a > 0$ with CIs [0.82, 0.93], [0.90, 1.05], and [0.99, 1.13]). Similarly, for the Gaussian solution we have $\Delta R^2_{\text{train}}$ CIs [.006, .013], [.011, .021], [.018, .040], and ΔR^2_{test} CIs, [.004, .012], [.015, .025], [.022, .040]. Finally, the GAM model did not improve on the exponential model despite the additional degrees of freedom $\Delta R^2_{\text{train}}$ CIs, [−.003, .001], [−.005, .001], [−.012, −.001], and ΔR^2_{test} CIs, [−.002, .002], [−.003, .002], [−.012, −.001]).

As an additional control to ensure that the exponential relation does not arise merely from the MDS algorithm itself, we randomized

Table 1

Full List of Model Evaluation Metrics on the Similarity Tasks and Their 95% Confidence Intervals Based on Split-Half Bootstrap Over Trials With 100 Repetitions

Category	Model	R^2	δR^2	r_{md}	δr_{md}	r_{c}	δr_{c}
Animals	Exponential	.854	.005	.906	.002	.982	.01
Animals	Linear	.784	.006	.868	.003	.968	.033
Animals	Quadratic	.854	.005	.906	.002	.98	.008
Animals	Gaussian	.840	.006	.899	.003	.970	.007
Animals	GAM	.858	.005	.907	.003	.983	.008
Fruits	Exponential	.575	.014	.69	.009	.998	.022
Fruits	Linear	.441	.014	.615	.008	.933	.048
Fruits	Quadratic	.563	.015	.683	.009	.991	.021
Fruits	Gaussian	.557	.016	.680	.009	.982	.015
Fruits	GAM	.581	.014	.692	.009	.997	.019
Vegetables	Exponential	.645	.010	.75	.007	.99	.009
Vegetables	Linear	.509	.010	.679	.005	.902	.015
Vegetables	Quadratic	.633	.011	.743	.007	.983	.009
Vegetables	Gaussian	.624	.012	.739	.007	.977	.008
Vegetables	GAM	.649	.010	.752	.007	.992	.008

Note. The measures are: R^2 = coefficient of determination; r_{md} = model–data Pearson correlation; r_{c} = model–data correlation corrected for attenuation; GAM = generalized additive model. δ indicates 95% confidence error (i.e., $\delta X = 1.96 \cdot \sigma_X$ where σ_X is the standard deviation of X). See Method for full details.

Table 2
Full List of Model Evaluation Metrics on the Generalization Tasks and Their 95% Confidence Intervals Based on Split-Half Bootstrap Over Trials With 100 Repetitions

Category	Model	R^2	δR^2	r_{md}	δr_{md}	r_c	δr_c
Animals	Exponential	.522	.031	.655	.01	.951	.015
Animals	Linear	.500	.032	.641	.01	.946	.026
Animals	Quadratic	.521	.031	.656	.01	.95	.014
Animals	Gaussian	.513	.031	.654	.01	.944	.012
Animals	GAM	.523	.031	.656	.01	.951	.015
Fruits	Exponential	.320	.023	.459	.012	.875	.023
Fruits	Linear	.301	.022	.449	.012	.863	.028
Fruits	Quadratic	.316	.023	.458	.012	.872	.023
Fruits	Gaussian	.304	.023	.451	.012	.864	.023
Fruits	GAM	.322	.023	.46	.012	.879	.024
Vegetables	Exponential	.202	.016	.265	.016	.928	.047
Vegetables	Linear	.184	.014	.263	.016	.903	.048
Vegetables	Quadratic	.192	.015	.261	.016	.918	.048
Vegetables	Gaussian	.173	.015	.248	.016	.915	.056
Vegetables	GAM	.209	.016	.27	.016	.933	.045

Note. See Table 1 for definitions of the various evaluation metrics. GAM = generalized additive model.

our data by shuffling the entries of the similarity and generalization matrices and then recomputed the generalization gradients (since these matrices are symmetric with unit diagonal, the shuffling boils down to the upper triangle). The resulting gradients are shown in Figure A3. As can be seen, the concave patterns largely disappear (compared to original Figure A2), and indeed, the VE of the exponential model dropped drastically relative to the original baseline: for the original similarity data sets we have 85.4%, 57.5%, and 64.5% VE for animals, fruits, and vegetables, respectively, whereas for the shuffled data sets we only have 16.0%, 19.2%, and 18.9%. Likewise, for the original generalization data sets, we have 52.2%, 32.0%, and 20.2% VE for animals, fruits, and vegetables, respectively, whereas for the shuffled data, we have 12.6%, 12.4%, and 11.9%. Moreover, the other models performed equally poorly on the shuffled data (<20% VE) suggesting that the relationship in this case is largely unstructured (see also Appendix C for an additional control regarding model misspecification). Viewed together, these results provide strong evidence for the universal law of generalization.

Discussion

Shepard's universal law of generalization stands out as a theoretical claim about cognition in its intended scope, covering all intelligent entities and all stimuli. However, previous work had only evaluated it using relatively small, simple sets of stimuli. We assessed its performance in large sets of high-dimensional natural images comprising more than 600,000 human judgments. Our results provide robust evidence for the validity of the universal law and extend its long research tradition into rich naturalistic domains. By analyzing both similarity and generalization judgments, we also confirmed that generalization over blank properties and default similarity judgments indeed capture shared sources of variance even when the dimensionality of the space is particularly large.

There are a number of limitations of the present work that can be further addressed by future research. First, our population was limited to online U.S. participants to allow for efficient scaling of data crowdsourcing. However, cross-cultural research is necessary in order to evaluate the extent to which our findings generalize beyond U.S. populations and English speakers (Blasi et al., 2022). Nevertheless, the fact that we focused specifically on widespread natural categories should facilitate such an investigation. Second, in the present work, we restricted ourselves to the visual modality, but one could equally consider natural categories in other primary modalities like the auditory and audio-visual (e.g., environmental sounds and scenes). While perhaps not as common as images, large behavioral data sets over such domains are becoming increasingly more accessible due to the growing interest in multimodal models in the machine learning community (see, e.g., Gemmeke et al., 2017; Marjieh et al., 2023). Third, future work could explore how the results of our generalization analysis vary when other blank properties are considered. Indeed, one might expect that different blank properties may activate different forms of inductive reasoning (Kemp & Tenenbaum, 2009) as well as intersubject variation. The extent to which these too support the universal law of generalization is an open question that requires further investigation. Finally, naturalistic stimuli provide much more space than artificial stimuli for interrogating the relationship between generalization and similarity. Our results showed a significant correlation between similarity and generalization, but it varied significantly across domains. This raises questions such as what features of a complex stimulus people rely on when generalizing from one stimulus to another, and how their weights differ when people evaluate similarity. This is particularly relevant when one considers the research prospects that are enabled by modern deep learning methods beyond traditional MDS. For example, one could use the framework of contrastive learning to incorporate similarity judgments in the training of deep networks (Muttenthaler et al., 2023) and then compare the generalization behavior of those networks against their learned representations and human data. In fact, this idea of learning an embedding based on one type of data (e.g., generalization judgments) and testing it on the other (e.g., similarity judgments) might be informative even within the framework of MDS. As a proof-of-concept, we computed MDS distances based on generalization data and plotted them against similarity data. The results again

revealed tight exponential relations (Figure C1), however, curiously the curve for the animal domain exhibited a small nonconcavity at small MDS distances which suggests a nontrivial relation between similarity and generalization judgments. One possible interpretation is that people are more conservative when they reason about scientific facts (e.g., having an enzyme) in a familiar domain like that of animals (relative to the less constrained similarity judgments). We hope to engage with these questions more systematically in future work.

More broadly, our work showcases the prospects of scaling up psychological research, providing unprecedented precision for tests of foundational hypotheses in cognitive science, as well as new avenues for exploration of naturalistic stimuli. If our goal is to identify universal psychological principles underlying human cognition, being able to test those principles in naturalistic settings is essential to making strong claims about their universality. Finding that the universal law of generalization holds for natural images provides support for its use as a component of other cognitive models applied to these rich and complex stimuli (e.g., Battleday et al., 2020; Sanders & Nosofsky, 2020), laying the groundwork for more extensive deployment and testing of models of human behavior based on psychological theory.

Constraints on Generality

As noted in the Discussion, we focused on online U.S. populations to allow for efficient and large-scale crowdsourcing on AMT. Further cross-cultural research is necessary in order to assess whether our findings generalize beyond U.S. populations and English speakers (Blasi et al., 2022). However, the fact that we focused specifically on natural categories that are widespread across cultures should facilitate such an investigation.

References

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, *11*(1), Article 5148. https://doi.org/10.1038/s41467-020-18946-z

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Chater, N., & Vitányi, P. M. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*(3), 346–369. https://doi.org/10.1016/S0022-2496(03)00013-0

Cheng, K. (2000). Shepard's universal law supported by honeybees in spatial generalization. *Psychological Science*, *11*(5), 403–408. https://doi.org/10.1111/1467-9280.00278

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017, March 5–9). *Audio set: An ontology and human-labeled dataset for audio events* [Paper presentation]. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, United States (pp. 776–780).

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*(1), 15–36. https://doi.org/10.1006/anbe.2003.2174

Harrison, P., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., Larrouy-Maestri, P., & Jacoby, N. (2020). Gibbs sampling with people. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 10659–10671). Curran Associates.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185. https://doi.org/10.1038/s41562-020-00951-3

Jensen, A. R. (1998). *The g Factor*. Praeger.

Jha, A., Peterson, J. C., & Griffiths, T. L. (2023). Extracting low-dimensional psychological representations from convolutional neural networks. *Cognitive Science*, *47*(1), Article e13226. https://doi.org/10.1111/cogs.13226

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58. https://doi.org/10.1037/a0014282

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27. https://doi.org/10.1007/BF02289565

Marjieh, R., Harrison, P. M., Lee, H., Deligiannaki, F., & Jacoby, N. (2022). *Reshaping musical consonance with timbral manipulations and massive online experiments*. bioRxiv. https://doi.org/10.1101/2022.06.14.496070

Marjieh, R., Van Rijn, P., Sucholutsky, I., Sumers, T., Lee, H., Griffiths, T. L., & Jacoby, N. (2023, May 1–5). *Words are all you need? Language as an approximation for human similarity judgments* [Paper presentation]. The Eleventh International Conference on Learning Representations, Kigali, Rwanda.

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001, July). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001* (Vol. 2, pp. 416–423). IEEE.

Muttenthaler, L., Linhardt, L., Dippel, J., Vandermeulen, R. A., Hermann, K., Lampinen, A. K., & Kornblith, S. (2023). *Improving neural network representations using human similarity judgments*. arXiv preprint arXiv:2306.04507. https://doi.org/10.48550/arXiv.2306.04507

Myung, I. J., & Shepard, R. N. (1996). Maximum entropy inference and stimulus generalization. *Journal of Mathematical Psychology*, *40*(4), 342–347. https://doi.org/10.1006/jmps.1996.0033

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*(2), 185–200. https://doi.org/10.1037/0033-295X.97.2.185

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Peer, M., Brunec, I. K., Newcombe, N. S., & Epstein, R. A. (2021). Structuring knowledge with cognitive maps and cognitive graphs. *Trends in Cognitive Sciences*, *25*(1), 37–54. https://doi.org/10.1016/j.tics.2020.10.004

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669. https://doi.org/10.1111/cogs.12670

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214. https://doi.org/10.1126/science.abe2629

Peterson, J. C., & Griffiths, T. L. (2017). *Evidence for the size principle in semantic and perceptual domains*. arXiv preprint arXiv:1705.03260. https://doi.org/10.48550/arXiv.1705.03260

Roads, B. D., & Love, B. C. (2021, June 20–25). *Enriching ImageNet with human similarity judgments and psychological embeddings* [Paper

presentation]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, United States (pp. 3547–3557).

Sanders, C. A., & Nosofsky, R. M. (2020). Training deep networks to construct a psychological feature space for a natural-object category domain. *Computational Brain & Behavior*, *3*, 229–251. https://doi.org/10.1007/s42113-020-00073-z

Servén, D., & Brummitt, C. (2018). *pyGAM: Generalized additive models in Python*. Zenodo.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140. https://doi.org/10.1007/BF02289630

Shepard, R. N. (1966). Metric structures in ordinal data. *Journal of Mathematical Psychology*, *3*(2), 287–315. https://doi.org/10.1016/0022-2496(66)90017-4

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398. https://doi.org/10.1126/science.210.4468.390

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656. https://doi.org/10.1126/science.aaq1118

Steyvers, M. (2006). Multidimensional scaling. In L. Nadel (Ed.), *Encyclopedia of cognitive science*. John Wiley & Sons. https://doi.org/10.1002/0470018860.s00585

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. https://doi.org/10.1017/S0140525X01000061

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Tversky, A., & Hutchinson, J. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review*, *93*(1), 3–22. https://doi.org/10.1037/0033-295X.93.1.3

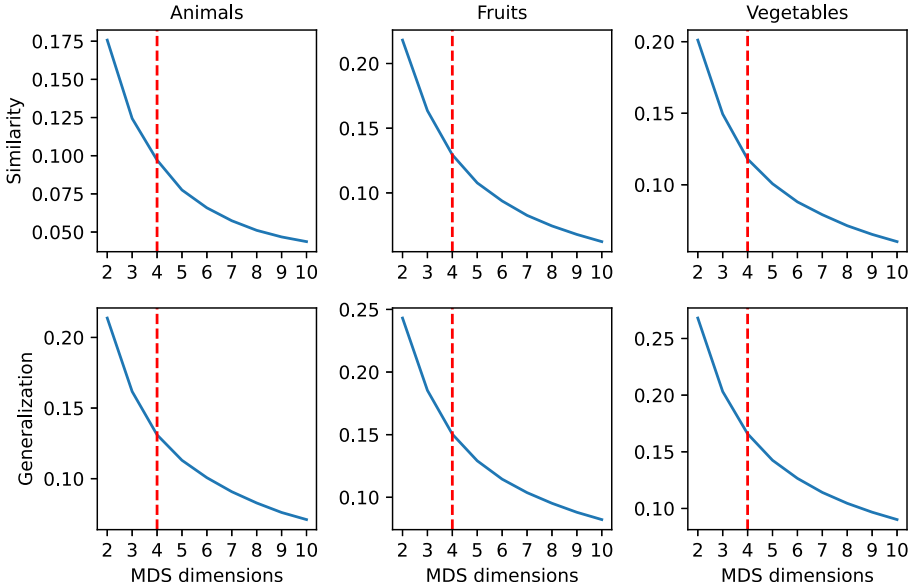
(*Appendices follow*)

Appendix A

Supplementary Information

Figure A1

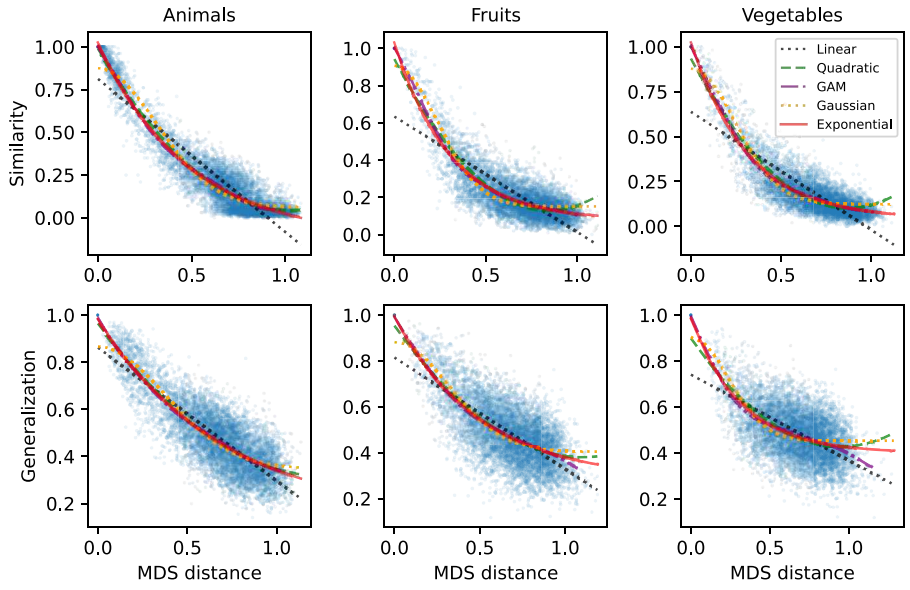
Nonmetric MDS Stress Curves for the Various Domains Considered in the Present Work as a Function of MDS Dimensions



Note. Based on this analysis, we selected the first dimensionality ($d = 4$, marked in dashed red vertical lines) for which all stress values across all data sets dropped below 0.2 (a standard threshold above which MDS fit is deemed poor; Kruskal, 1964). MDS = multidimensional scaling. See the online article for the color version of this figure.

Figure A2

Raw Generalization Gradients Across Domains of Natural Images and Tasks With the Optimal Fitted Models Overlaid

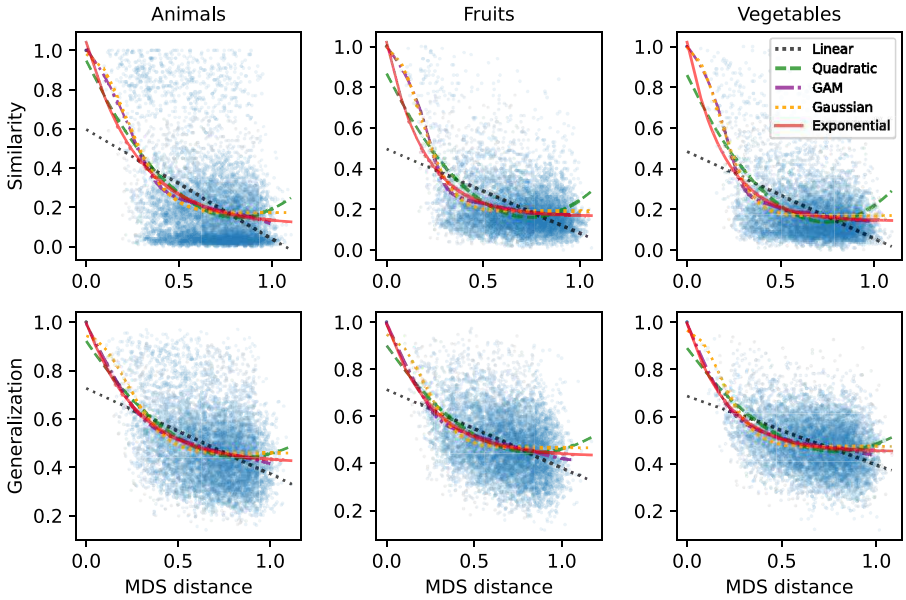


Note. GAM = generalized additive model; MDS = multidimensional scaling. See the online article for the color version of this figure.

(*Appendices continue*)

Figure A3

Raw Generalization Gradients for Randomly Shuffled Similarity and Generalization Matrices Across Domains of Natural Images and Tasks With the Optimal Fitted Models Overlaid



Note. GAM = generalized additive model; MDS = multidimensional scaling. See the online article for the color version of this figure.

Table A1

Exponential Model Parameters of the Form $g(x) = ae^{-bx} + c$

Task	Category	*a*	δ*a*	*b*	δ*b*	*c*	δ*c*
Similarity	Animals	1.179	0.009	2.022	0.043	−0.138	0.009
Similarity	Fruits	0.969	0.008	3.243	0.075	0.072	0.005
Similarity	Vegetables	1.005	0.007	3.08	0.056	0.032	0.005
Generalization	Animals	0.844	0.018	1.525	0.085	0.154	0.022
Generalization	Fruits	0.711	0.012	2.029	0.114	0.28	0.016
Generalization	Vegetables	0.604	0.009	2.916	0.144	0.378	0.01

Note. Errors (δ) indicate 95% confidence intervals (i.e., $\delta X = 1.96 \cdot \sigma_X$ where σ_X is the standard deviation of *X*) bootstrapped over trials with 100 repetitions.

Table A2

Linear Model Parameters of the Form $g(x) = ax + b$

Task	Category	*a*	δ*a*	*b*	δ*b*
Similarity	Animals	−0.917	0.005	0.827	0.003
Similarity	Fruits	−0.629	0.008	0.639	0.006
Similarity	Vegetables	−0.671	0.007	0.646	0.005
Generalization	Animals	−0.568	0.007	0.86	0.005
Generalization	Fruits	−0.484	0.008	0.807	0.006
Generalization	Vegetables	−0.39	0.009	0.745	0.006

Note. Errors (δ) indicate 95% confidence intervals (i.e., $\delta X = 1.96 \cdot \sigma_X$ where σ_X is the standard deviation of *X*) bootstrapped over trials with 100 repetitions.

(*Appendices continue*)

Table A3
Quadratic Model Parameters of the Form $g(x) = ax^2 + bx + c$

Task	Category	a	δa	b	δb	c	δc
Similarity	Animals	0.906	0.021	−1.872	0.023	1.007	0.005
Similarity	Fruits	1.152	0.03	−1.951	0.033	0.957	0.009
Similarity	Vegetables	1.134	0.024	−1.968	0.026	0.956	0.007
Generalization	Animals	0.436	0.027	−1.069	0.03	0.977	0.007
Generalization	Fruits	0.488	0.036	−1.058	0.04	0.951	0.01
Generalization	Vegetables	0.528	0.035	−1.017	0.039	0.906	0.01

Note. Errors (δ) indicate 95% confidence intervals (i.e., $\delta X = 1.96 \cdot \sigma_X$ where σ_X is the standard deviation of X) bootstrapped over trials with 100 repetitions.

Table A4
Gaussian Model Parameters of the Form $g(x) = ae^{-bx^2} + c$

Task	Category	a	δa	b	δb	c	δc
Similarity	Animals	0.824	0.006	5.058	0.148	0.058	0.004
Similarity	Fruits	0.767	0.010	9.055	0.350	0.151	0.003
Similarity	Vegetables	0.776	0.009	8.325	0.256	0.121	0.002
Generalization	Animals	0.525	0.006	3.802	0.211	0.351	0.007
Generalization	Fruits	0.481	0.010	5.109	0.386	0.400	0.007
Generalization	Vegetables	0.458	0.014	7.981	0.777	0.440	0.006[a]

Note. Errors (δ) indicate 95% confidence intervals (i.e., $\delta X = 1.96 \cdot \sigma_X$ where σ_X is the standard deviation of X) bootstrapped over trials with 100 repetitions.
[a] In two out of the 100 runs, the optimizer did not converge for the generalization/vegetables condition, so we excluded those runs.

Table A5
Complementary List of Evaluation Measures on the Similarity Tasks and Their 95% Confidence Intervals Based on Split-Half Bootstrap Over Trials With 100 Repetitions

Category	Model	r_{dd}	δr_{dd}	r_{mm}	δr_{mm}	r_{md}	δr_{md}	ΔBIC	δΔBIC
Animals	Exponential	.890	.004	.956	.019	.906	.002	0.0	0.0
Animals	Linear	.890	.004	.905	.063	.868	.003	2782.613	138.579
Animals	Quadratic	.890	.004	.960	.016	.906	.002	−4.387	35.575
Animals	Gaussian	.890	.004	.965	.014	.899	.003	654.683	119.148
Animals	GAM	.890	.004	.958	.016	.907	.003	−19.403	103.188
Fruits	Exponential	.591	.011	.808	.041	.690	.009	0.0	0.0
Fruits	Linear	.591	.011	.737	.083	.615	.008	1970.43	139.787
Fruits	Quadratic	.591	.011	.803	.039	.683	.009	207.94	56.923
Fruits	Gaussian	.591	.011	.812	.031	.680	.009	293.346	77.199
Fruits	GAM	.591	.011	.815	.036	.692	.009	58.402	39.297
Vegetables	Exponential	.640	.010	.897	.018	.750	.007	0.0	0.0
Vegetables	Linear	.640	.010	.884	.033	.679	.005	2321.842	136.625
Vegetables	Quadratic	.640	.010	.893	.019	.743	.007	236.349	43.698
Vegetables	Gaussian	.640	.010	.893	.016	.739	.007	416.224	80.967
Vegetables	GAM	.640	.010	.898	.017	.752	.007	69.554	39.957

Note. The measures are: r_{dd} = data–data Pearson correlation; r_{mm} = model–model Pearson correlation; r_{md} = model–data Pearson correlation; $\Delta BIC = BIC_{model} - BIC_{exponential}$ the BIC relative to the exponential model in each category and $BIC = n \log(RSS/n) + k \log n$ where $RSS = \sum_i (x_i - \hat{x}_i)^2$ is the residual sum of squares between data and model, n is the number of data points and k is the number of fitted parameters. δ indicates 95% confidence error (i.e., $\delta X = 1.96 \cdot \sigma_X$ where σ_X is the standard deviation of X). BIC = Bayesian information criterion; GAM = generalized additive model; RSS = residual sum of squares.

(*Appendices continue*)

Table A6

Complementary List of Evaluation Measures on the Generalization Tasks and Their 95% Confidence Intervals Based on Split-Half Bootstrap Over Trials With 100 Repetitions

Category	Model	r_{dd}	δr_{dd}	r_{mm}	δr_{mm}	r_{md}	δr_{md}	ΔBIC	δΔBIC
Animals	Exponential	.577	.011	.823	.037	.655	.010	0.0	0.0
Animals	Linear	.577	.011	.796	.056	.641	.010	320.879	44.845
Animals	Quadratic	.577	.011	.825	.036	.656	.010	10.57	6.792
Animals	Gaussian	.577	.011	.832	.033	.654	.010	139.584	30.441
Animals	GAM	.577	.011	.824	.037	.656	.010	148.101	14.716
Fruits	Exponential	.417	.013	.661	.041	.459	.012	0.0	0.0
Fruits	Linear	.417	.013	.649	.049	.449	.012	188.706	53.238
Fruits	Quadratic	.417	.013	.660	.04	.458	.012	36.861	8.414
Fruits	Gaussian	.417	.013	.652	.039	.451	.012	170.940	26.997
Fruits	GAM	.417	.013	.658	.042	.46	.012	135.5	15.998
Vegetables	Exponential	.188	.016	.437	.053	.265	.016	0.0	0.0
Vegetables	Linear	.188	.016	.454	.061	.263	.016	158.541	55.12
Vegetables	Quadratic	.188	.016	.432	.055	.261	.016	88.4	20.665
Vegetables	Gaussian	.188	.016	.391	.052	.248	.016	257.959	51.938
Vegetables	GAM	.188	.016	.447	.054	.27	.016	101.458	26.378

Note. See Table A5 for definition of the various metrics. BIC = Bayesian information criterion; GAM = generalized additive model.

Appendix B

Binned Analysis

Table B1

Exponential Model Parameters of the Form $g(x) = ae^{-bx} + c$

Task	Category	*a*	δ*a*	*b*	δ*b*	*c*	δ*c*
Similarity	Animals	1.154	0.017	2.194	0.101	−0.097	0.021
Similarity	Fruits	1.059	0.028	3.396	0.14	0.071	0.01
Similarity	Vegetables	1.069	0.022	3.195	0.129	0.032	0.009
Generalization	Animals	0.844	0.044	1.525	0.201	0.153	0.056
Generalization	Fruits	0.75	0.044	1.769	0.347	0.229	0.068
Generalization	Vegetables	0.608	0.036	2.546	0.661	0.344	0.063

Note. Errors (δ) indicate 95% confidence intervals (i.e., $\delta X = 1.96 \cdot \sigma_X$ where σ_X is the standard deviation of *X*) bootstrapped over trials with 100 repetitions and $n = 100$ bins.

Table B2

Linear Model Parameters of the Form $g(x) = ax + b$

Task	Category	*a*	δ*a*	*b*	δ*b*
Similarity	Animals	−0.882	0.046	0.817	0.018
Similarity	Fruits	−0.723	0.03	0.721	0.016
Similarity	Vegetables	−0.77	0.032	0.731	0.015
Generalization	Animals	−0.57	0.02	0.869	0.01
Generalization	Fruits	−0.491	0.02	0.819	0.012
Generalization	Vegetables	−0.385	0.03	0.747	0.016

Note. Errors (δ) indicate 95% confidence intervals (i.e., $\delta X = 1.96 \cdot \sigma_X$ where σ_X is the standard deviation of *X*) bootstrapped over trials with 100 repetitions and $n = 100$ bins.

(*Appendices continue*)

Table B3
Quadratic Model Parameters of the Form $g(x) = ax^2 + bx + c$

Task	Category	a	δa	b	δb	c	δc
Similarity	Animals	0.958	0.033	−1.919	0.033	1.008	0.008
Similarity	Fruits	1.203	0.086	−2.057	0.093	0.992	0.024
Similarity	Vegetables	1.189	0.077	−2.067	0.081	0.984	0.019
Generalization	Animals	0.43	0.072	−1.063	0.073	0.974	0.015
Generalization	Fruits	0.386	0.099	−0.96	0.11	0.93	0.026
Generalization	Vegetables	0.37	0.172	−0.841	0.189	0.858	0.043

Note. Errors (δ) indicate 95% confidence intervals (i.e., $\delta X = 1.96 \cdot \sigma_X$ where σ_X is the standard deviation of X) bootstrapped over trials with 100 repetitions and $n = 100$ bins.

Table B4
Gaussian Model Parameters of the Form $g(x) = ae^{-bx^2} + c$

Task	Category	a	δa	b	δb	c	δc
Similarity	Animals	0.801	0.009	5.808	0.282	0.071	0.008
Similarity	Fruits	0.748	0.024	9.581	0.633	0.150	0.006
Similarity	Vegetables	0.755	0.018	8.863	0.531	0.120	0.006
Generalization	Animals	0.522	0.012	3.796	0.361	0.343	0.016[a]
Generalization	Fruits	0.467	0.015	3.810	0.789	0.359	0.028
Generalization	Vegetables	0.388	0.032	5.336	2.224	0.405	0.041

Note. Errors (δ) indicate 95% confidence intervals (i.e., $\delta X = 1.96 \cdot \sigma_X$ where σ_X is the standard deviation of X) bootstrapped over trials with 100 repetitions and $n = 100$ bins.
[a] In one out of the 100 runs, the optimizer did not converge for the generalization/animal condition, so we excluded that run.

Table B5
Full List of Model Evaluation Metrics on the Similarity Tasks and Their 95% Confidence Intervals Based on Split-Half Bootstrap Over Trials With 100 Repetitions and n = 100 Bins

Category	Model	R^2	δR^2	r_{md}	δr_{md}	r_c	δr_c
Animals	Exponential	.994	.002	.997	.001	.999	.001
Animals	Linear	.911	.014	.954	.005	.956	.005
Animals	Quadratic	.993	.002	.996	.001	.998	.001
Animals	Gaussian	.975	.004	.987	.002	.989	.001
Animals	GAM	.997	.001	.998	.001	1.0	.0
Fruits	Exponential	.986	.005	.993	.002	.998	.002
Fruits	Linear	.812	.018	.902	.005	.906	.006
Fruits	Quadratic	.969	.008	.984	.003	.989	.003
Fruits	Gaussian	.972	.007	.986	.002	.991	.002
Fruits	GAM	.991	.005	.994	.002	1.0	.001
Vegetables	Exponential	.991	.004	.995	.001	.999	.001
Vegetables	Linear	.837	.014	.915	.004	.919	.004
Vegetables	Quadratic	.979	.005	.989	.002	.993	.002
Vegetables	Gaussian	.978	.006	.989	.002	.993	.002
Vegetables	GAM	.993	.004	.996	.002	1.0	.001

Note. The measures are: R^2 = coefficient of determination; r_{md} = model–data Pearson correlation; r_c = model–data correlation corrected for attenuation; GAM = generalized additive model. δ indicates 95% confidence error (i.e., $\delta X = 1.96 \cdot \sigma_X$ where σ_X is the standard deviation of X). See Method for full details.

(*Appendices continue*)

Table B6

Full List of Model Evaluation Metrics on the Generalization Tasks and Their 95% Confidence Intervals Based on Split-Half Bootstrap Over Trials With 100 Repetitions and n = 100 Bins

Category	Model	R^2	δR^2	r_{md}	δr_{md}	r_c	δr_c
Animals	Exponential	.989	.01	.994	.004	1.0	.003
Animals	Linear	.951	.02	.975	.007	.98	.007
Animals	Quadratic	.989	.008	.994	.003	.999	.003
Animals	Gaussian	.978	.025	.989	.008	.995	.004
Animals	GAM	.992	.004	.993	.006	1.0	.001
Fruits	Exponential	.97	.016	.985	.006	.997	.006
Fruits	Linear	.932	.02	.966	.006	.978	.007
Fruits	Quadratic	.965	.017	.982	.007	.994	.006
Fruits	Gaussian	.948	.020	.973	.008	.986	.008
Fruits	GAM	.982	.01	.986	.009	1.0	.003
Vegetables	Exponential	.935	.052	.965	.021	.991	.017
Vegetables	Linear	.888	.051	.945	.018	.97	.02
Vegetables	Quadratic	.924	.044	.958	.023	.985	.016
Vegetables	Gaussian	.892	.055	.943	.024	.972	.020
Vegetables	GAM	.974	.015	.964	.039	1.0	.005

Note. See Table 1 for definitions of the various evaluation metrics. GAM = generalized additive model.

Table B7

Complementary List of Evaluation Measures on the Similarity Tasks and Their 95% Confidence Intervals Based on Split-Half Bootstrap Over Trials With 100 Repetitions and n = 100 Bins

Category	Model	r_{dd}	δr_{dd}	r_{mm}	δr_{mm}	r_{md}	δr_{md}	ΔBIC	$\delta\Delta$BIC
Animals	Exponential	.996	.002	1.0	.0	.997	.001	0.0	0.0
Animals	Linear	.996	.002	1.0	.0	.954	.005	257.735	27.564
Animals	Quadratic	.996	.002	1.0	.0	.996	.001	16.567	20.269
Animals	Gaussian	.996	.002	1.0	.0	.987	.002	141.725	24.090
Animals	GAM	.996	.002	.999	.001	.998	.001	21.103	40.222
Fruits	Exponential	.99	.005	1.0	.0	.993	.002	0.0	0.0
Fruits	Linear	.99	.005	1.0	.0	.902	.005	226.335	31.753
Fruits	Quadratic	.99	.005	1.0	.0	.984	.003	71.777	25.337
Fruits	Gaussian	.990	.005	1.0	.0	.986	.002	61.495	28.891
Fruits	GAM	.99	.005	.999	.001	.994	.002	39.543	30.356
Vegetables	Exponential	.992	.004	1.0	.0	.995	.001	0.0	0.0
Vegetables	Linear	.992	.004	1.0	.0	.915	.004	251.392	35.803
Vegetables	Quadratic	.992	.004	1.0	.0	.989	.002	72.835	31.745
Vegetables	Gaussian	.992	.004	1.0	.0	.989	.002	77.422	37.945
Vegetables	GAM	.992	.004	.999	.001	.996	.002	52.324	28.149

Note. The measures are: r_{dd} = data–data Pearson correlation; r_{mm} = model–model Pearson correlation; r_{md} = model–data Pearson correlation; $\Delta\text{BIC} = \text{BIC}_{\text{model}} - \text{BIC}_{\text{exponential}}$ the BIC relative to the exponential model in each category and $\text{BIC} = n \log(\text{RSS}/n) + k \log n$ where $\text{RSS} = \sum_i (x_i - \hat{x}_i)^2$ is the residual sum of squares between data and model, n is the number of data points and k is the number of fitted parameters. δ indicates 95% confidence error (i.e., $\delta X = 1.96 \cdot \sigma_X$ where σ_X is the standard deviation of X). BIC = Bayesian information criterion; GAM = generalized additive model; RSS = residual sum of squares.

(*Appendices continue*)

Table B8

Complementary List of Evaluation Measures on the Generalization Tasks and Their 95% Confidence Intervals Based on Split-Half Bootstrap Over Trials With 100 Repetitions and $n = 100$ Bins

Category	Model	r_{dd}	δr_{dd}	r_{mm}	δr_{mm}	r_{md}	δr_{md}	ΔBIC	$\delta\Delta$BIC
Animals	Exponential	.989	.007	1.0	.001	.994	.004	0.0	0.0
Animals	Linear	.989	.007	1.0	.0	.975	.007	125.869	39.802
Animals	Quadratic	.989	.007	1.0	.001	.994	.003	3.683	16.468
Animals	Gaussian	.989	.007	.999	.019	.989	.008	59.403	46.175
Animals	GAM	.989	.007	.998	.006	.993	.006	53.514	53.277
Fruits	Exponential	.976	.014	1.0	.001	.985	.006	0.0	0.0
Fruits	Linear	.976	.014	1.0	.0	.966	.006	66.156	44.537
Fruits	Quadratic	.976	.014	.999	.002	.982	.007	13.808	9.359
Fruits	Gaussian	.976	.014	.999	.004	.973	.008	48.669	17.033
Fruits	GAM	.976	.014	.996	.008	.986	.009	40.117	46.13
Vegetables	Exponential	.950	.042	.998	.004	.965	.021	0.0	0.0
Vegetables	Linear	.950	.042	1.0	.0	.945	.018	39.1	54.258
Vegetables	Quadratic	.950	.042	.995	.011	.958	.023	12.989	18.99
Vegetables	Gaussian	.950	.042	.992	.020	.943	.024	40.484	22.805
Vegetables	GAM	.950	.042	.979	.037	.964	.039	9.844	63.035

Note. See Table B7 for definition of the various metrics. BIC = Bayesian information criterion; GAM = generalized additive model.

Appendix C

Additional Controls

Controlling for Model Misspecification

To ensure that our similarity and generalization matrices are consistent with low-dimensional spatial representations irrespective of any embedding procedure, we computed two diagnostic measures proposed by Tversky and Hutchinson (1986). These measures are known as centrality and reciprocity and they characterize the nearest neighbor statistics induced by a given proximity matrix s_{ij}. Centrality is defined as $C = \frac{1}{n}\sum_i N_i^2$ where N_i is the number of stimuli whose nearest neighbor is stimulus i, and n is the overall number of stimuli. Likewise, reciprocity is defined as $R = \frac{1}{n}\sum_i R_i$ where R_i is the rank of stimulus i in the proximity order of its nearest neighbor (ties are broken at random). Tversky and Hutchinson (1986) compellingly argued that low-dimensional spatial representations tend to generically produce low R, C scores (theoretically < 2, but in practice < 3), whereas nonspatial conceptual representations with semantic hubs (e.g., superordinate categories that are nearest neighbors to many other concepts) tend to produce high R, C values (in some cases as high as 10–15). We computed these measures on our data sets and found that they are indeed small as summarized in Table C1.

Table C1

Centrality (C) and Reciprocity (R) Measures for the Different Behavioral Matrices

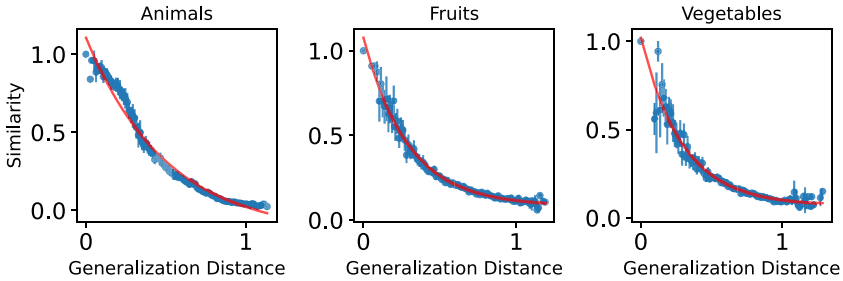
Task	Category	C	R
Similarity	Animals	1.82	2.14
Similarity	Fruits	1.78	1.77
Similarity	Vegetables	1.80	1.98
Generalization	Animals	1.70	2.02
Generalization	Fruits	1.87	1.99
Generalization	Vegetables	1.72	2.21

(*Appendices continue*)

Hybrid Generalization Gradients

Figure C1

Hybrid Generalization Gradients Across Domains of Natural Images and Tasks With the Optimal Exponential Model Overlaid



Note. MDS distances were computed based on the generalization data and plotted against similarity scores. Error bars indicate 95% confidence intervals. MDS = multidimensional scaling. See the online article for the color version of this figure.

Received June 28, 2023
Revision received September 23, 2023
Accepted November 11, 2023 ■