Psychonomic Bulletin & Review (2020) 27:791–800
https://doi.org/10.3758/s13423-020-01754-3

BRIEF REPORT

Perceived similarity ratings predict generalization success after traditional category learning and a new paired-associate learning task

Stefania R. Ashby[1] · Caitlin R. Bowman[1] · Dagmar Zeithamova[1]

Published online: 29 May 2020

Abstract
The current study investigated category learning across two experiments using face-blend stimuli that formed face families controlled for within- and between-category similarity. Experiment 1 was a traditional feedback-based category-learning task, with three family names serving as category labels. In Experiment 2, the shared family name was encountered in the context of a face-full name paired-associate learning task, with a unique first name for each face. A subsequent test that required participants to categorize new faces from each family showed successful generalization in both experiments. Furthermore, perceived similarity ratings for pairs of faces were collected before and after learning, prior to generalization test. In Experiment 1, similarity ratings increased for faces within a family and decreased for faces that were physically similar but belonged to different families. In Experiment 2, overall similarity ratings decreased after learning, driven primarily by decreases for physically similar faces from different families. The post-learning category bias in similarity ratings was predictive of subsequent generalization success in both experiments. The results indicate that individuals formed generalizable category knowledge prior to an explicit demand to generalize and did so both when attention was directed towards category-relevant features (Experiment 1) and when attention was directed towards individuating faces within a family (Experiment 2). The results tie together research on category learning and categorical perception and extend them beyond a traditional category-learning task.

Keywords Category learning · Perceived similarity · Memory generalization

Introduction

Categorization helps us organize information from the world around us into meaningful clusters relevant to behavior. A hallmark of category knowledge is the ability to categorize new instances (memory generalization), allowing us to use our prior experiences to guide decisions in novel situations (Knowlton & Squire, 1993; R.M. Nosofsky & Zaki, 1998; Poldrack et al., 2001; Reber, Stark, & Squire, 1998). Category knowledge also results in biases in perception, which can manifest as increased perceived similarity of items within a category, decreased perceived similarity of items from different categories, or a combination of both (Beale & Keil, 1995; Goldstone, 1994a; Goldstone, Lippa, & Shiffrin, 2001; Kurtz, 1996; Livingston, Andrews, & Harnad, 1998). These perceptual biases are often thought to reflect stretching of the perceptual space along the category-relevant dimensions and/or shrinking along the category-irrelevant dimension, resulting from shifts of attention to the relevant features (Goldstone & Steyvers, 2001; Kruschke, 1996; Medin & Schaffer, 1978; Nosofsky, 1991; Nosofsky, 1986). While a category bias on peception can emerge relatively quickly following category learning, it remains unknown to what degree it reflects the quality of category knowledge and relates to subsequent categorization and generalization performance. If category learning results in changes of perceptual space and persistent attentional shifts to category-relevant features, the degree of category bias on perception should be a good indicator of the quality of category knowledge. On the other hand, if good learners more accurately encode all information – which may allow them to better determine which information is category relevant and which irrelevant – then the degree of

✉ Dagmar Zeithamova
dasa@uoregon.edu

[1] Department of Psychology, University of Oregon, 1227, Eugene, OR 97403, USA

category bias may not be a good predictor of category knowledge. Thus, one goal of the current study was to measure both category bias in perception and generalization in a single study to determine to what degree category bias in perception following category learning can be used as a measure of generalizable category knowledge by predicting performance on unstudied items.

Most categorization studies explicitly instruct participants to learn categories. Several studies have also compared categorization tasks that focus on contrast across categories and commonalities within categories to identification tasks that focus on learning stimulus-specific information (Nosofsky, 1986; Shepard & Chang, 1963; Shepard, Hovland, & Jenkins, 1961). However, in the real world, category information can be available alongside information about specific items or individuals, without an explicit goal to form category knowledge. For example, when attending a wedding and meeting many new individuals, one's objective is to remember individual people and learn their unique names. Yet, some guests may share last names, providing an opportunity to also extract categorical structure across individuals. Past work has shown that category knowledge can be extracted without explicit instruction (Aizenstein et al., 2000; Bozoki, Grossman, & Smith, 2006; Gabay, Dick, Zevin, & Holt, 2015; Kéri, Kálmán, Kelemen, Benedek, & Janka, 2001; Love, 2002; Reber, Gitelman, Parrish, & Mesulam, 2003; Wattenmaker, 1993). However, how category learning proceeds when category information is available but instructions emphasize learning of specific information is rarely addressed. While some show that categorization performance can be predicted from performance on identification tasks that emphasize discrimination of individual items (Nosofsky, 1986), others have found that learning and generalizing concept information is more challenging when learning is focused on discrimination of individual stimuli (Soto & Wasserman, 2010). Thus, in Experiment 2, our goal was to assess signatures of category knowledge – generalization and category bias on perception – in a task that emphasizes memory for stimulus-specific information and more closely resembles an episodic paired-associate learning task than a traditional category-learning task.

In the current paper, we assessed (1) category bias on perception, (2) category generalization success, and (3) their relationship after traditional category learning (Experiment 1) and after a novel task where category information was available but instructions emphasized stimulus-specific information (Experiment 2). Participants were shown faces that belonged to three categories (families), designated by a family name. Face stimuli were created as blends of never-seen "parent" faces, resulting in increased physical similarity between faces that shared a parent. Some physically similar faces were members of the same family while others were members of different families, allowing us to dissociate the effect of category membership from physical similarity. In Experiment 1, faces were encountered in the context of a traditional feedback-based category learning task, emphasizing similarities among faces belonging to the same family and how they contrast with faces belonging to different families. In Experiment 2, faces were encoded through observational, face-full name paired-associate learning. While family names were identical to Experiment 1, with each family name shared across several faces, first names were unique for each face, requiring participants to remember individual faces and differentiate faces within each family. Perceived similarity ratings were collected immediately before and after learning to test for the emergence of category bias in perception. We also tested participants' ability to generalize family names to new face-blend stimuli. The category bias in perceived similarity ratings after learning was related to subsequent generalization success in order to determine the extent to which category bias in perception reflects the quality of category knowledge.

The current design allowed us to also address additional questions regarding the nature of category bias in perception. First, what drives category bias in perception has been variable across studies. Some studies have shown *between-category expansion* or *acquired distinctinctiveness,* where items across a learned category boundary become more discriminable (Beale & Keil, 1995; Folstein, Palmeri, & Gauthier, 2013; Goldstone, 1994a; Gureckis & Goldstone, 2008; Wallraven, Bülthoff, Waterkamp, van Dam, & Gaißert, 2014), and are perceived as more dissimilar after category learning (Goldstone et al., 2001). Category bias can also manifest as *within-category compression* or *acquired equivalence*, where items within a learned category become less discriminable (Gureckis & Goldstone, 2008; Soto, 2019) and are perceived as more similar after category learning (Goldstone et al., 2001; Kurtz, 1996; Livingston et al., 1998). As relatively few studies show both compression and expansion effects following category learning (but see Goldstone et al., 2001; Gureckis & Goldstone, 2008), we were interested to what degree both expansion and compression effects can be observed after category learning of the face-blend stimuli with equated within-category and between-category physical similarity. Furthermore, the aforementioned studies on learning-related category bias have focused on traditional category learning. Thus, the degree to which within-category compression and between-category expansion can be observed after learning that emphasizes memory for stimulus-specific information remains unknown.

Finally, using perceived similarity to probe category knowledge in Experiment 2 can help us link research on the emergence of conceptual knowledge to another area of generalization research: episodic inference. Episodic inference refers to the ability to integrate information across distinct experiences that share content to infer new information (e.g., infering that two people are likely a couple after seeing each of them with the same child on different occasions). Whether people spontaneously integrate memories of related events as they are encoded (Cai et al., 2016; Gershman, Schapiro, Hupbach, & Norman, 2013; Schlichting, Mumford, & Preston, 2015; Shohamy & Wagner, 2008; Zeithamova, Dominick, & Preston, 2012) or whether links between related memories are formed in response to generalization demands (Banino, Koster, Hassabis, & Kumaran, 2016; Carpenter & Schacter, 2017, 2018) remains debated. Here, observing evidence for the formation of a category representation under conditions that minimize generalization demands – such as observing category bias in perceived similarity ratings after learning but before the explicit generalization test —would suggest that participants may extract category information and form category representations spontaneously.

Method

Participants

Healthy participants – N = 39 in Experiment 1 and N = 43 in Experiment 2 – were recruited from the University of Oregon community via the university SONA research system and received course credit for their participation. Except for the learning phase, all procedures were identical across experiments and are presented together. All participants provided written informed consent, and experimental procedures were approved by Research Compliance Services at the University of Oregon. From Experiment 1, four participants were excluded due to chance performance (accuracy ≤ .33) in categorizing the training faces. From Experiment 2, participants were excluded for failing to make responses on more than 25% of categorization trials (n = 3) and incomplete data (n = 1). After exclusions, analyses were carried out with the remaining 35 participants for Experiment 1 (M_{age} = 20.43, SD_{age} = 2.58, 18–32 years, 21 females) and 39 participants for Experiment 2 (M_{age} = 19.26, SD_{age} = 1.13, 18–23 years, 21 females). These sample sizes provide 80% power for detecting medium-size effects (d ≥ 0.5) using planned one-sample and paired t-tests and strong (r ≥ .5) correlations, as determined in G-Power (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007).

Stimuli

Stimuli were grayscale images of blended faces constructed by morphing two unaltered face images together using FantaMorph Version 5 by Abrosoft. We used blended faces because it allowed us to maintain realistic-looking stimuli while also controlling for within- and across-category physical similarity. Faces were also convenient for creating the face-name learning task in Experiment 2 that was intuitive for the participants and yielded the right level of difficulty as verified through a pilot study. Prior work has shown that category effects differ based on whether morphed faces are constructed from parents within one race versus across two races (Levin & Angelone, 2002). Thus, we restricted all parent faces to be Caucasian to ensure that the resulting face-blend stimuli were comparably similar to all other faces with a shared parent. Additionally, all parent faces were of a single gender (male) to ensure that face-blends maintained a realistic appearance. Parent faces were compiled over several years from multiple sources, including the Dallas Face Database (O'Toole et al., 2005), CVL Face Database provided by the Computer Vision Laboratory, University of Ljubljana, Slovenia (Peer, 1999), and Google Image Search. Faces were selected primarily based on whether they would blend well with other faces (e.g., visibility of both ears, no facial hair, etc.) but were not formally equated for features such as attractiveness or memorability.

The stimulus structure is presented in Fig. 1. For each participant, three category-relevant parent faces and three category-irrelevant parent faces were randomly selected from a total set of 20 faces. Each of the three category-relevant parent faces were individually morphed with each of the three category-irrelevant parent faces with equal weight given to each parent face (50/50 blend). The resulting nine blended faces were then used as training stimuli. Faces that shared a category-relevant parent shared a family name (belonged to the same category). Faces that shared a category-irrelevant parent belonged to different families. As faces sharing any parent (category-relevant or category-irrelevant) shared physical traits, physical similarity alone was not diagnostic of category membership. Because of the blending procedure used, an equal number of category-relevant and category-irrelevant parent faces were selected to provide equal exposure to the relevant and irrelevant category features. With an uneven number of relevant versus irrelevant parent faces (e.g., two relevant parent faces blended with multiple irrelevant parent faces to create family members), unsupervised learning could take place, making the features of the relevant parent faces more prominent through increased exposure instead of being category-learning driven. We chose a three-way category structure, which provided nine blended faces to learn and therefore 36 pairwise similarity rating comparisons. We determined that the three-way structure provided the best balance of a reasonable number of training stimuli to learn but still provided adequate pairwise comparisons for similarity ratings. Generalization stimuli were new faces created by blending category-relevant parent faces with 14 remaining parent faces not used for creation of the training faces.

Procedure

Both experiments consisted of the following phases: passive viewing, pre-learning similarity ratings, learning (different in

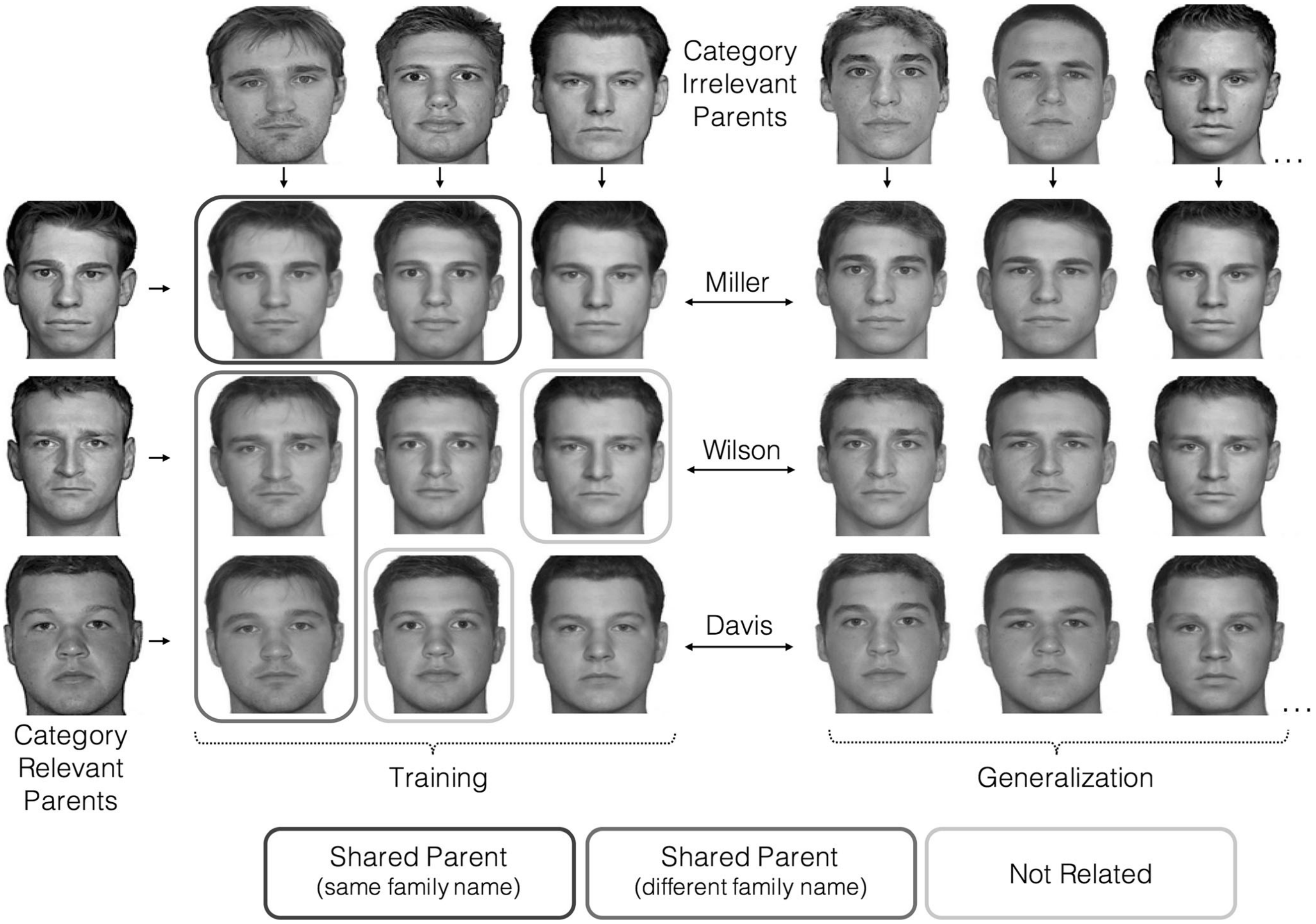
Category Irrelevant Parents
Miller
Wilson
Davis
Category Relevant Parents
Training
Generalization
Shared Parent (same family name)
Shared Parent (different family name)
Not Related

Fig. 1 Example of face-blend stimuli. Parent faces on the leftmost side are designated "category relevant parents" as these parents determined family membership – Miller, Wilson, or Davis – during learning and generalization. Parent faces across the top are designated "category-irrelevant parents" as these parents introduced physical similarity across families but did not determine categories. Three category-irrelevant parents were used for learning. The rightmost three category-irrelevant parents are a subset of new faces used for generalization. Parent faces were never viewed by participants, only the resulting blended faces. The face-blending procedure produced pairs of faces that shared a category-relevant parent and belonged to the same family (shared parent – same family name; example indicated with dark grey box), pairs of faces that shared a category-irrelevant parent and belonged to different families (shared parent – different family name; example indicated with medium grey box). Non-adjacent pairs did not share a parent and were not related (example indicated with light grey boxes)

each Experiment), passive viewing, post-learning similarity ratings, and category generalization. Additionally, Experiment 2 included cued-recall of face-name associations before the category-generalization phase. Self-paced breaks separated the phases.

Passive viewing To familiarize participants with the stimuli and give them an idea of the degree of similarity between all faces before collecting perceived similarity ratings, participants first viewed each of the nine training stimuli individually, once in a random order without any labels and without making any responses. Face-blends were shown for 3 s with a 1-s inter-stimulus-interval (ISI). Passive viewing of the face-blends immediately before the pre- and post-learning similarity rating phases was also included as a pilot of a future neuroimaging experiment. No responses were collected during viewing.

Pre-learning similarity ratings To validate that participants were sensitive to the similarity structure among faces introduced by the blending process and to obtain baseline similarity ratings, participants rated the subjective similarity of pairs of faces to be used during the learning phase. All possible 36 pairwise comparisons of the nine training faces were presented and participants rated the similarity of the two faces on a scale from one to six (1 = two faces appeared very dissimilar, 6 = two faces appeared very similar). Face pairs and the similarity rating scale were displayed for 5 s with a 1-s ISI. Face pairs were then binned into three conditions for analyses depending on whether they (1) shared a parent and a family name, (2) shared a parent face but did not share a family name, or (3) did not share a parent face (see example pairs in Fig. 1).

Learning phase *Experiment 1: Feedback-based category learning.* On each trial, a training face was presented on the screen along with family names (Miller, Wilson, Davis) as response options. Participants were instructed to indicate family membership via a button press and received corrective feedback after each trial. Each face was viewed simultaneously with the family name response options on the screen for 4 s, received corrective feedback for 1 s, and trials were separated by a 1-s ISI. Each face was presented 16 times total, evenly split across two blocks.

Experiment 2: Observational learning of face – full name associations. To test the robustness of category learning outside of a traditional categorization task, Experiment 2 provided an opportunity to form associations between faces from the same families in the context of a face-full name associative learning task. On each trial, participants studied a face-name pair that was presented on-screen for 2 s and then made a prospective memory judgement for 2 s on a scale from 1 to 4 (1 = definitely will not remember, 4 = definitely will remember). Trials were separated by a 4-s ISI and participants viewed each face-name pair 12 times, evenly split across three blocks. Prospective memory judgments were included to facilitate participant engagement with the observational learning task and were not considered further. Family names were identical to Experiment 1 and shared across faces whereas first names were unique to each face. While the inclusion of face-specific first names required participants to differentiate individual faces, the inclusion of the shared family names provided an opportunity to form links between related faces. We designed the task to determine to what degree experiences that overlap in content (here, last name) tend to affect perception and be related in memory, bridging traditional category research with research on generalization through episodic inference (Schlichting & Preston, 2015; Zeithamova et al., 2012). However, we subsequently discovered similarities between our task and a study by Medin, Dewey, and Murphy, (1983). In Medin et al. (1983), participants also learned first and shared last names of faces but under a feedback-based categorization paradigm rather than a paired-associate observational paradigm. Because our task did not employ feedback-based learning, participants were not provided with cues as to the number of first names or surnames. The fact that family names were repeated across faces or that there was a category structure among faces was not explicitly emphasized to participants. This allowed us to see if we could replicate results from Experiment 1 under very different conditions, in a task that does not resemble traditional category learning and where category information is present but not emphasized.

Post-learning similarity ratings Perceived similarity ratings were repeated after the learning phase with the same timing as pre-learning ratings. Of main interest was a potential category bias in perceived similarity, i.e., whether faces that shared a parent would be rated as more similar when they had the same family name than when they had different family names.

Cued recall of face-name associations Experiment 2 included a self-paced cued-recall task of face-name associations. Participants viewed each training face individually on a computer screen and handwrote the full name of each face on a sheet of paper. Participants advanced the trials at their own pace but were not able to skip faces or go back and look at faces already named. Participants were encouraged to make their best guess as to the first and family names of each face even if they were not confident in their memory.

Generalization phase As the last phase of both Experiments, category knowledge was tested directly using categorization of old and new faces. In addition to the nine training faces, participants categorized 42 never-seen faces, consisting of 14 new blends of each of the three category-relevant parent faces. Participants were asked to select via button press the family name for each face, which were presented individually for 4 s, from the three options (Miller, Wilson, Davis) presented on the screen. Trials were separated by an 8-s ISI. No feedback was provided, and participants were encouraged to make their best guess when unsure of family membership.

Results

Learning phase

Experiment 1: Feedback-based category learning Overall percent correct across training was 76% (SD = 14%), which was well above chance (33% for three categories; one-sample t(34) = 17.66, p < .001, d = 3.01). Categorization accuracy improved across training, from 66% in the first half to 85% in the second half (t(34) = 9.72, p < .001, d = 1.63), demonstrating learning over time.

Experiment 2: Observational learning of full name – face associations Observational learning provided no measure of accuracy from the learning phase. Therefore, in Experiment 2 a cued-recall task was included to assess how well participants learned the face-full name pairs. Participants recalled on average 52% of first names and 65% of family names.

Similarity ratings

We compared mean face similarity ratings in each pair-type (shared parent-same family name, shared parent-different family name, not related) using repeated-measures ANOVA. Analyses were performed separately in each phase (pre-learning, post-learning). We also assessed learning-related rating changes by comparing ratings across phases. For all ANOVAs, a Greenhouse-Geisser correction for degrees of freedom (denoted as *GG*) was used wherever Mauchly's test indicated a violation of the assumption of sphericity.

Experiment 1 Pre-learning ratings (Fig. 2A) demonstrated that participants were sensitive to the physical similarity structure introduced with the face-blending procedure. A one-way, repeated-measures ANOVA showed a significant effect of pair type (F(2, 68) = 58.74, p < .001, η_p^2 = .63), driven by lower perceived similarity for faces that did not share a parent compared to those that did share a parent (with or without shared family name, both t > 9.17, p < .001, d > 1.50). Faces that shared a parent were perceived as equally similar to one another irrespective of whether they also shared the same – not yet presented – family name (t(34) = -0.17, p = .87, d = 0.03).

Post-learning ratings (Fig. 2B) revealed a category bias on perceived similarity: pairs of faces sharing a parent and family name were perceived as significantly more similar than faces that shared a parent but not a family name (M_{diff} = 0.72, SD_{diff} = 1.41, t(34) = 3.02, p = .005, d = 0.51). Faces that shared a parent remained rated as more similar than unrelated faces (both t > 6.85, p < .001, d > 1.15).

To further test the effect of learning, we conducted a 2 × 3 (timepoint [pre-learning, post-learning] × pair-type [shared parent-same family name, shared parent-different family name, not related]) repeated-measures ANOVA. There was no main effect of timepoint (F(1, 34) = 0.04, p = .85, η_p^2 = .001). There was a significant main effect of pair-type (F(1.63, 55.38) = 61.21, p < .001, η_p^2 = .64, *GG*), and a significant interaction between timepoint and pair-type (F(1.64, 55.88) = 11.85, p < .001, η_p^2 = .25, *GG*). Follow-up pre-post comparisons within each pair-type (Fig. 2C) revealed that this interaction was driven by both a significant *increase* in similarity ratings for faces sharing a parent and a family name (t(34) = 3.02, p = .005, d = 0.51) and a significant *decrease* in

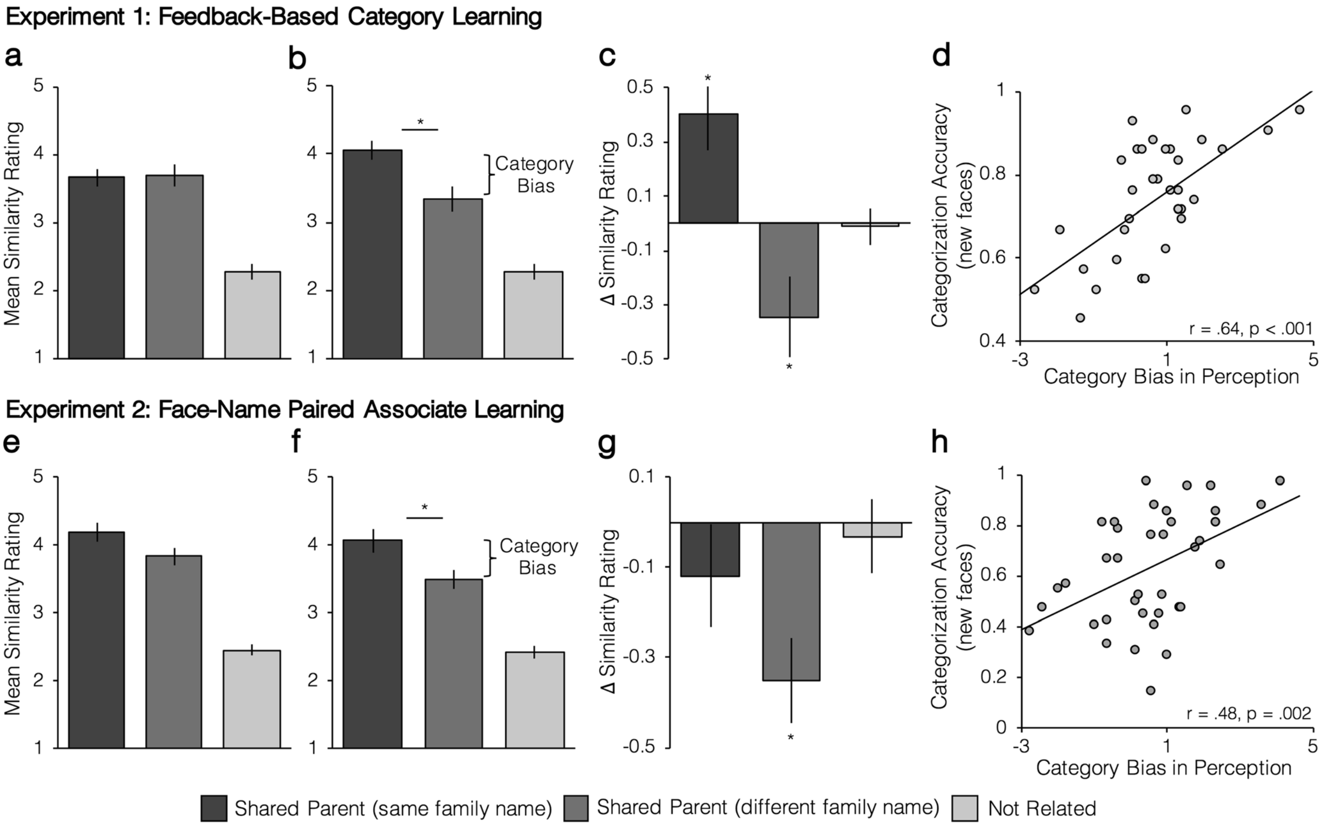

Fig. 2 Top panel shows results from the traditional category learning experiment. Bottom panel shows results from the face-name paired associate-learning experiment. **A and E.** Average similarity ratings for faces that share a parent and family name, faces that only share a parent, and faces that don't share any parents before learning. **B and F.** Average similarity ratings for the same pairwise comparisons after learning. Asterisk represents a significant (p < .05) difference in post-learning similarity ratings for faces that belong to the same family vs. faces that share physical similarity but belong to different families (i.e., a category bias in perception). **C and G.** Changes in similarity ratings from pre- to post-learning. Asterisk denotes significant (p < .05) increases and decreases in perceived similarity for faces. **D and H.** Positive relationship between indirect (category bias in perception) and direct (categorization accuracy for new faces) measures of memory generalization

similarity ratings for faces only sharing a parent but not a family name (t(34) = -2.33, p = .026, d = -0.39). There was no significant change in similarity ratings for faces that did not share a parent (t(34) = -0.18, p = .86, d = -0.03).

Experiment 2 As in Experiment 1, participants were sensitive to the face similarity structure. Pre-learning similarity ratings (Fig. 2E) differed significantly among pair types (F(1.46, 55.47) = 72.22, p < .001, η_p^2 = .655, *GG*), driven by lower perceived similarity of faces that did not share a parent compared to faces that shared a parent (with and without shared family names, both t > 10.65, p < .001, d > 1.70). For faces that shared a parent, ratings did not significantly differ when face pairs had the same or different – not yet presented – family names (t(38) = 1.82, p = .077, d = 0.29). A category bias was found in post-learning ratings (Fig. 2F) with pairs of faces sharing a parent and family name perceived as significantly more similar than faces that shared a parent but not a family name (M_{diff} = 0.58, SD_{diff} = 1.52; t(38) = 2.39, p = .022, d = 0.38).

Testing the effect of learning, the 2 × 3 (timepoint × pair-type) repeated-measures ANOVA revealed a significant main effect of timepoint (F(1, 38) = 5.20, p = .028, η_p^2 = .120), with overall similarity ratings being lower post-learning than pre-learning (M_{pre} = 3.49, SD_{pre} = 0.51; M_{post} = 3.33, SD_{post} = 0.59; t(38) = -2.28, p = .028, d = 0.37). There was also a significant main effect of pair-type (F(1.28, 48.60) = 60.42, p < .001, η_p^2 = .614, *GG*), and a significant interaction between timepoint and pair-type (F(1.67, 63.37) = 4.21, p = .03, η_p^2 = .10, *GG*). Follow-up pre-post comparisons within each pair-type (Fig. 2G) revealed that the interaction was driven by a significant *decrease* in similarity ratings for faces sharing a parent but not a family name (t(38) = -3.71, p = .001, d = -0.59), but there were no significant changes in similarity ratings for other pair-types (both t < -1.04, p > .30, d < -0.18). Thus, changes in perceived similarity were affected by category membership in both experiments.

Although not significant (p = .077), we noted a numerical tendency towards a category bias in pre-learning similarity ratings. Parent faces were randomly selected for each participant to serve as category-relevant or category-irrelevant parents, but some of the category-relevant parent faces may have been more salient, leading to a numerically greater pre-learning similarity rating. Thus, we tested whether the post-learning category bias on perceived similarity was reliably greater than pre-learning bias. A 2 × 2 (timepoint [pre-learning, post-learning] × pair-type [shared parent-same family name, shared parent-different family name]) repeated-measures ANOVA showed only a marginal interaction between timepoint and condition (F(1, 38) = 2.87, p = .098, η_p^2 = .07). We thus controlled for pre-learning similarity rating differences in subsequent analyses that assessed the relationship of post-learning ratings and generalization performance.

Category generalization

Experiment 1 Participants correctly categorized 85% of training faces (SD = 17%) and 74% of new faces (SD = 13%), which was well above chance (33% for three categories; both one-sample t(34) > 18.12, p < .001, d > 3.06). A paired-samples t-test showed higher categorization accuracy for the training faces than for the new faces (t(34) = 5.48, p < .001 , d = 0.93). We next tested whether the category bias on perceived similarity ratings (an indirect measure of category knowledge) was related to subsequent generalization success. A Pearson correlation showed a significant positive relationship between the category bias on perceived similarity ratings and generalization accuracy (r(33) = .64, p < .001; Fig. 2D). The category bias on perceived similarity in the post-learning phase was a significant predictor of subsequent generalization performance even when pre-learning similarity ratings were considered (multiple regression: pre-learning differences in perceived similarity β = .30, t(34) = 1.80, p = .08; post-learning category bias β = .46, t(34) = 2.75, p = .01).

Experiment 2 Participants correctly categorized 70% of training faces (SD = 23%) and 64% of new faces (SD = 22%), which was well above chance (33% for three categories; both one-sample t(38) > 8.65, p < .001, d > 1.38). A paired-samples t-test showed higher categorization accuracy for the training faces than for new faces (t(38) = 2.12, p = .04, d = 0.34). The post-learning category bias on perceived similarity ratings was significantly correlated with generalization accuracy (Pearson's r(37) = .48, p = .002; Fig. 2H). Further, the category bias was a significant predictor of subsequent generalization performance even when pre-learning similarity ratings were controlled for (multiple regression: pre-learning category bias β = -.22, t(38) = -0.86, p = .40; post-learning category bias β = .66, t(38) = 2.57, p = .01).

Discussion

The current study investigated category learning using measures of perceived similarity and category generalization across two experiments. Face-blend stimuli were used to control physical similarity within and across categories (families). Experiment 1 was a traditional feedback-based category-learning task, with three family names serving as category labels. In Experiment 2, the shared family name category label was encountered in the context of a face-full name paired-associate learning task, where first names were unique for each face. Participants were able to successfully apply category labels to new faces in both experiments, demonstrating that category information can be extracted in support of generalization even when task goals do not emphasize learning categories at encoding. Past work of incidental category learning has shown that individuals can extract category structures when not instructed using patterns of physical similarity as category cues (Aizenstein et al., 2000; Love, 2002; Reber et al., 2003; Wattenmaker, 1993). We extend these prior findings by showing that category structure can also be extracted when category membership is dissociable from physical similarity and further when individuals are actively learning information that differentiates individual items *within* the same category.

Learning-related changes in perceived similarity ratings were observed in both experiments. In both cases, following learning, participants rated faces sharing a category label as more similar than equally physically similar faces that did not share a category label. These results extend prior studies finding changes in perceived similarity as a result of explicit category learning (Goldstone, 1994b, 1994a; Livingston et al., 1998) to a novel task that exposes participants to a category label but requires individuation of stimuli within a category. Observing category bias after the face-name paired-associate learning also indicates that the mere presence of a shared piece of information can bias perception even outside the context of a traditional category-learning task.

The current results also indicate that similarity ratings provide a useful tool to index category knowledge while minimizing explicit generalization demands. In both experiments, category bias in similarity ratings observed after learning predicted subsequent generalization of category information to new examples. To our knowledge, this is the first study relating the strength of a perceptual category bias to the quality of learned category information (as measured by generalization success). The finding that good category generalizers were those who showed the greatest distortion in perceptual representations (rather than those with representations better aligned with physical similarity) is consistent with the view that category bias in perception results from learning-related attentional shifts and differential weighting of perceptual features based on their category relevance (Goldstone & Steyvers, 2001; Kruschke, 1996; Medin & Schaffer, 1978; Nosofsky, 1991; Nosofsky, 1986). Our findings tie together research on categorical perception and concept generalization, and newly indicate that perceived similarity ratings reflect the quality of new category knowledge robustly across two distinct tasks involving category learning.

Interestingly, while perceptual biases occurred in both experiments, they took different forms. In Experiment 1, similarity ratings for faces within a family increased while similarity ratings for faces that were physically similar but belonged to different families decreased. These results provide a new example of a category structure in which both within-category compression and between-category expansion are observed after traditional feedback-based category learning (Gurekis & Goldstone, 2008; Goldstone, Lippa & Shiffrin, 2001), and aligns well with the task demands of treating some stimuli as distinct and some as equivalent. Based on prior work on attentional shifts after category learning (Goldstone & Steyvers, 2001; Kruschke, 1996; Nosofsky, 1991), this result indicates that participants both focused more strongly on features that differentiate between categories (features of the relevant parent faces) and decreased attention to features that do not differentiate between categories (features of the irrelevant parent faces that affected physical similarity of faces but not family membership).

In contrast, the changes in perceived similarity after the face-name paired-associate learning in Experiment 2 were primarily driven by decreased similarity for faces that were physically similar but belonged to different families. We did not observe increases in perceived similarity ratings for faces belonging to the same family. While more difficult category structures are thought to trigger within-category compression (Pothos & Reppa, 2014), this does not explain differences observed here as category structure was the same across experiments and category learning was easier rather than more difficult in Experiment 1, where compression was observed. Rather, we suspect that learning goals at encoding drove the differences in the pattern of category bias between experiments. Although it is not possible to rule out a contribution from other factors, such as feedback-based versus observational learning, the goal of learning a full name for each face (including the unique first names) in the paired-associates task was likely a key factor. It required participants to look for differences between *all* faces, even faces within the same family, in order to differentiate between categories as well as between "brothers" within the same family. That meant that all features remained relevant for task goals in Experiment 2, as the features of category-irrelevant parent faces were important for discriminating two members of the same family, such as differentiating Brad Miller from Ryan Miller. Thus, participants could not simply ignore the category-irrelevant features as they could in Experiment 1.

Notably, the category bias was measured *after* learning but *before* the explicit generalization test, meaning that the category bias was present prior to explicit generalization demands. Yet, the presence of a shared piece of information (same last name) was sufficient to affect how faces became represented, even in Experiment 2 where no features were irrelevant for the task at hand. This finding is consistent with the notion that people spontaneously link related episodes into an integrated representation at encoding (Shohamy & Wagner, 2008; Zeithamova, Dominick, & Preston, 2012) rather than in response to explicit generalization demands (Banino et al., 2016; Carpenter & Schacter, 2017, 2018). As a strategic decision to rate faces with the same last name as more similar can contribute to biases in similarity ratings (Goldstone, 1994b; Goldstone et al., 2001), we cannot definitively attribute our findings to spontaneous integration during learning. However, our results *do* indicate that evidence for the formation of category knowledge can be demonstrated even when generalization task demands are greatly minimized, and outside of a traditional category learning task. The nature of the resulting category representations – such as whether they are exemplar-based (Hintzman, 1986; Medin & Schaffer, 1978), prototype-based (Homa, Cross, Cornell, Goldman, & Shwartz, 1973; Posner & Keele, 1968), or cluster-based (Love & Medin, 1998) – cannot be resolved in the current study as any model of category learning that postulates learning-related attentional shifts would predict the emergence of perceptual biases.

In summary, we build on long lines of research on category learning (for reviews, see Ashby & Maddox, 2011; Seger, 2008) and categorical perception (for reviews, see Goldstone & Hendrickson, 2010; Harnad, 2006) by demonstrating that category bias in perception reflects the quality of learned category knowledge. We further extend prior work beyond traditional category learning, to demonstrate perceptual biases and successful generalization even after learning that emphasizes individuation of category members, with the specific pattern of learning-related perceptual shifts reflecting goals during learning. Lastly, relating our results to hypotheses generated from studies of episodic inference, our data align with the notion that individuals may spontaneously link related information at encoding, prior to explicit demands to generalize.

Author Note This work was supported by the National Institute on Aging Grant F32-AG-054204 awarded to Caitlin R. Bowman.

Open Practices None of the experiments discussed in the current report were preregistered. Data and materials for all experiments are freely available in the *Blended-Face Similarity Ratings and Categorization Tasks* repository on the Open Science Framework (https://osf.io/e8htb/).

References

Aizenstein, H., MacDonald, A., Stenger, V., Nebes, R., Larson, J., Ursu, S., & Carter, C. (2000). Complementary category learning systems identified using fMRI. *Journal of Cognitive Neuroscience*, *12*(6), 977–987.

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0 *Annals of the New York Academy of Sciences*, *1224*(1), 147–161. https://doi.org/10.1111/j.1749-6632.2010.05874.x

Banino, A., Koster, R., Hassabis, D., & Kumaran, D. (2016). Retrieval-based model accounts for striking profile of episodic memory and generalization. *Scientific Reports*, *6*, 1–15. https://doi.org/10.1038/srep31330

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, *57*, 217–239.

Bozoki, A., Grossman, M., & Smith, E. E. (2006). Can patients with Alzheimer's disease learn a category implicitly? *Neuropsychologia*, *44*(5), 816–827. https://doi.org/10.1016/j.neuropsychologia.2005.08.001

Cai, D. J., Aharoni, D., Shuman, T., Shobe, J., Biane, J., Song, W., … Silva, A. J. (2016). A shared neural ensemble links distinct contextual memories encoded close in time. *Nature*, *534*(7605), 115–118. https://doi.org/10.1038/nature17955

Carpenter, A. C., & Schacter, D. L. (2017). Flexible retrieval: When true inferences produce false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(3), 335–349.

Carpenter, A. C., & Schacter, D. L. (2018). False memories, false preferences: Flexible retrieval mechanisms supporting successful inference bias novel decisions. *Journal of Experimental Psychology: General*, *147*(7), 988–1004. https://doi.org/10.1037/xge0000391

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, *23*(4), 814–823. https://doi.org/10.1093/cercor/bhs067

Gabay, Y., Dick, F. K., Zevin, J. D., & Holt, L. L. (2015). Incidental auditory category learning. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(4), 1124–1138. https://doi.org/10.1037/xhp0000073

Gershman, S. J., Schapiro, A. C., Hupbach, A., & Norman, K. A. (2013). Neural context reinstatement predicts memory misattribution. *Journal of Neuroscience*, *33*(20), 8590–8595. https://doi.org/10.1523/JNEUROSCI.0096-13.2013

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

Goldstone, R. L. (1994b). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*, 125–157.

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

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, *130*(1), 116–139.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43. https://doi.org/10.1016/S0010-0277(00)00099-8

Gureckis, T. M., & Goldstone, R. L. (2008). The Effect of the Internal Structure of Categories on Perception. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. https://doi.org/10.4314/jlt.v44i2.71793

Harnad, S. (2006). Categorical Perception. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science* (pp. 1–5). https://doi.org/10.1002/0470018860.s00490

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review*, *93*(4), 411–428.

Homa, D., Cross, J., Cornell, D., Goldman, D., & Shwartz, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology*, *101*(1), 116–122. https://doi.org/10.1037/h0035772

Kéri, S., Kálmán, J., Kelemen, O., Benedek, G., & Janka, Z. (2001). Are Alzheimer's disease patients able to learn visual prototypes? *Neuropsychologia*, *39*(11), 1218–1223. https://doi.org/10.1016/S0028-3932(01)00046-X

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, *262*, 1747–1749. https://doi.org/10.1126/science.8259522

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*(2), 225–247. https://doi.org/10.1080/095400996116893

Kurtz, K. J. (1996). Category-based similarity. In G. W. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (p. 290).

Levin, D. T., & Angelone, B. L. (2002). Categorical perception of race. *Perception*, *31*(5), 567–578. https://doi.org/10.1068/p3315

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

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9*(4), 829–835.

Love, B. C., & Medin, D. L. (1998). SUSTAIN: A model of human category learning. *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, 671–676.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Medin, D. L., Dewey, G. I., & Murphy, T. D. (1983). Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 607–625. https://doi.org/10.1037/0278-7393.9.4.607

Nosofsky, R M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(1), 3–27. https://doi.org/10.1037/0096-1523.17.1.3

Nosofsky, R.M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals. *Psychological Science*, *9*(4), 247–255.

Nosofsky, Robert M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

O'Toole, A. J., Harms, J., Snow, S. L., Hurst, D. R., Pappas, M. R., Ayyad, J. H., & Abdi, H. (2005). A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(5), 812–816. https://doi.org/10.1109/TPAMI.2005.90

Peer, P. (1999). CVL Face Database. Retrieved from Computer Vision Lab, Faculty of Computer and Information Science, University of Ljubljana, Slovenia. website: http://www.lrv.fri.uni-li.si/facedb.html

Poldrack, R., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature*, *414*, 546–550. https://doi.org/10.1038/35107080

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363.

Pothos, E. M., & Reppa, I. (2014). The fickle nature of similarity change as a result of categorization. *Quarterly Journal of Experimental Psychology*, *67*(12), 2425–2438. https://doi.org/10.1080/17470218.2014.931977

Reber, P. J., Stark, C. E. L., & Squire, L. R. (1998). Contrasting cortical activity associated with category memory and recognition memory. *Learning & Memory*, *5*, 420–428. https://doi.org/10.1101/lm.5.6.420

Reber, P. J., Gitelman, D. R., Parrish, T. B., & Mesulam, M. M. (2003). Dissociating explicit and implicit category knowledge with fMRI. *Journal of Cognitive Neuroscience*, *15*(4), 574–583. https://doi.org/10.1162/089892903321662958

Schlichting, M. L., & Preston, A. R. (2015). Memory integration: Neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences*, *1*, 1–8. https://doi.org/10.1016/j.cobeha.2014.07.005

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications*, *6*, 1–10. https://doi.org/10.1038/ncomms9151

Seger, C. A. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neuroscience and Biobehavioral Reviews*, *32*(2), 265–278. https://doi.org/10.1016/j.neubiorev.2007.07.010

Shepard, R. N., & Chang, J. J. (1963). Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology*, *65*(1), 94–102. https://doi.org/10.1037/h0043732

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42. https://doi.org/10.1037/h0093825

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: Hippocampal-midbrain encoding of overlapping events. *Neuron*, *60*, 378–389. https://doi.org/10.1016/j.neuron.2008.09.023

Soto, F. A. (2019). Categorization training changes the visual representation of face identity. *Attention, Perception, and Psychophysics*, *81*(5), 1220–1227. https://doi.org/10.3758/s13414-019-01765-w

Soto, F. A., & Wasserman, E. A. (2010). Missing the Forest for the Trees: Object-discrimination Learning Blocks Categorization Learning. *Psychological Science*, *21*(10), 1510–1517. https://doi.org/10.1177/0956797610382125

Wallraven, C., Bülthoff, H. H., Waterkamp, S., van Dam, L., & Gaißert, N. (2014). The eyes grasp, the hands see: Metric category knowledge transfers between vision and touch. *Psychonomic Bulletin and Review*, *21*(4), 976–985. https://doi.org/10.3758/s13423-013-0563-4

Wattenmaker, W. D. (1993). Incidental concept learning, feature frequency, and correlated properties. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(1), 203–222. https://doi.org/10.1037/0278-7393.19.1.203

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron*, *75*(1), 168–179. https://doi.org/10.1016/j.neuron.2012.05.010

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.