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

Stefania R. Ashby, Caitlin R. Bowman, Dagmar Zeithamova

University of Oregon

Author Note

Stefania R. Ashby, Department of Psychology, University of Oregon; Caitlin R. Bowman, Department of Psychology, University of Oregon; Dagmar Zeithamova, Department of Psychology, University of Oregon.

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

Correspondence concerning this article should be addressed to Dagmar Zeithamova, Department of Psychology, 1227 University of Oregon, Eugene, OR 97403.

Email: dasa@uoregon.edu

Phone: 541-346-6731

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 similarities (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

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

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, allowing us to generalize prior knowledge and guide decisions in novel situations. Category-learning tasks have thus been widely used to study memory generalization (Knowlton & Squire, 1993; Nosofsky & Zaki, 1998; Poldrack et al., 2001; Reber, Stark, & Squire, 1998). Prior work has contended that memory generalization relies on memory representations that form during learning, linking information across related experiences (Knowlton & Squire, 1993; Schapiro, Turk-Browne, Botvinick, & Norman, 2017; Schlichting, Zeithamova, & Preston, 2014; Shohamy & Wagner, 2008; Zeithamova, Schlichting, & Preston, 2012). Other work maintains that specific memory traces are formed during learning and that generalization judgements may be computed from specific memories either on-the-fly at retrieval (Hintzman, 1984; Kruschke, 1992; Nosofsky, 1988) or by linking information across experiences in response to explicit generalization demands (Carpenter & Schacter, 2017, 2018; Squire, 1992; Teyler & DiScenna, 1986; Winocur, Moscovitch, & Sekeres, 2007). Finding ways to detect category knowledge in behavior outside of generalization demands will help us to determine whether or not people spontaneously link related experiences as they are encountered.

Category knowledge alters perception such that items learned to belong to the same category are perceived as more similar while items learned to belong to different categories are perceived as less similar after learning (Beale & Keil, 1995; Folstein, Palmeri, & Gauthier, 2013; Goldstone, 1994a; Goldstone, Lippa, & Shiffrin, 2001; Livingston, Andrews, & Harnad, 1998; Rosch & Mervis, 1975). Thus, measures of perceived similarity may be useful for assessing category knowledge without creating an explicit generalization demand. In the current report, we conducted two experiments testing whether measures of perceived similarity can reveal the formation of category knowledge prior to an explicit generalization test. 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 (Figure 1). 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. We tested participants’ ability to extract commonalities across faces belonging to the same family and generalize family names to new face-blend stimuli. We also measured category knowledge indirectly, using perceived similarity ratings immediately before and after learning to determine to what degree people link related experiences prior to explicit demands to generalize. The category bias in perceived similarity ratings after learning was related to subsequent generalization success to determine the utility of using perceived similarity ratings as a measure of category learning. The same measures were collected in Experiment 2, where 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 the participant to differentiate faces within each family. This allowed us to test to what degree the results from Experiment 1 replicate outside of a traditional category learning task context.

Methods

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 will be 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 $\leq$ .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 ($d \geq 0.5$) effects using planned one-sample and paired t-tests and strong ($r \geq .5$) 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. 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.

The stimulus structure is presented in Figure 1. For each participant, three category-relevant parent faces and three category-irrelevant parent faces were randomly selected from a total set of twenty 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

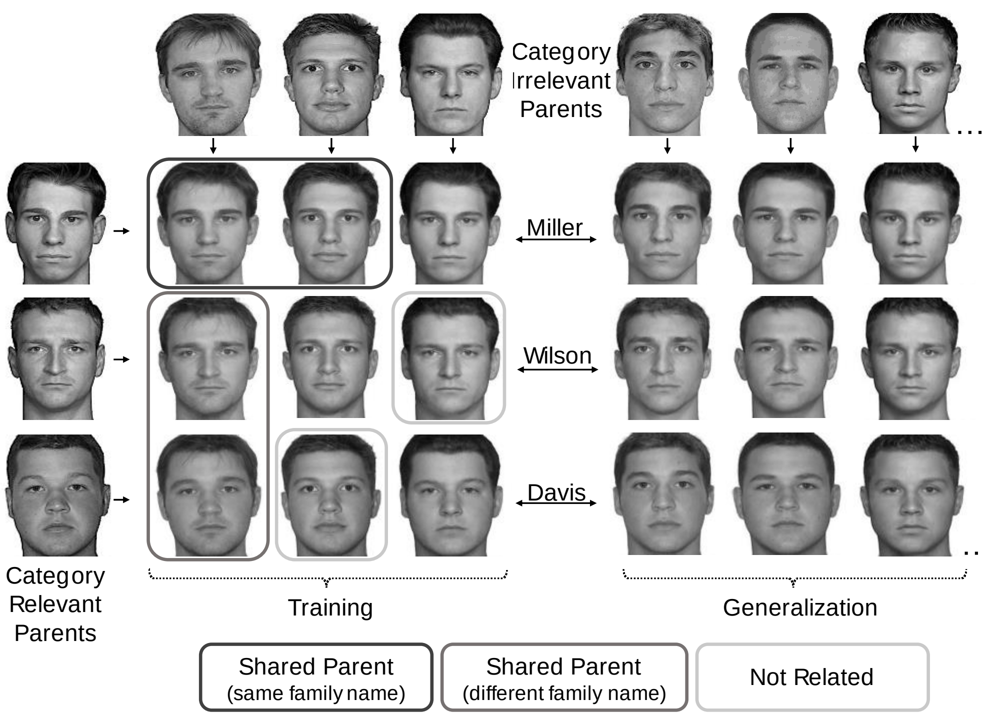

Figure 1. Example face-blend stimuli. Parent faces on the leftmost side are designated “category relevant parents” as these parents determined family membership—Miller, Wilson, or Davis—during learning and generalization. Parent faces across the top are designated “category-irrelevant parents” as these 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).

parent (category-relevant or category-irrelevant) shared physical traits, physical similarity alone was not diagnostic of category membership. Generalization stimuli were new faces created by blending category-relevant parent faces with fourteen 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 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 3s with a 1s 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 9 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 5s with a 1s 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 Figure 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 4s, received corrective feedback for 1s, and trials were separated by a 1s ISI. Each face was presented 16 times total, evenly split across 2 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 link 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 and then made a prospective memory judgement on a scale from one to four (1 = definitely will not remember, 4 = definitely will remember). 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. The fact that family names were repeated across faces or that there was a category structure among faces was not explicitly emphasized to participants. Each face-name pair was presented on screen for 2s after which the prospective memory judgment scale appeared beneath the face-name pair for an additional 2s. Trials were separated by a 4s ISI. Participants viewed each face-name pair twelve times, evenly split across 3 blocks.

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 4s, from the three options (Miller, Wilson, Davis) presented on the screen. Trials were separated by an 8s 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 ($F(2, 68) = 58.74$, $p < .001$, $\eta_p^2 = .63$), driven by lower perceived similarity for faces that did not share a parent compared to those that shared 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 x 3 (timepoint [pre-learning, post-learning] x 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 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 x 3 (timepoint x pair-type) repeated-measures ANOVA revealed a significant main effect of timepoint ($F(1, 38) = 5.20$, $p = .028$, $\eta_p^2 = .120$), 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 ($F(1.28, 48.60) = 60.42$, $p < .001$, $\eta_p^2 = .614$, *GG*), and a significant interaction between timepoint and pair-type ($F(1.67, 63.37) = 4.21$, $p = .03$, $\eta_p^2 = .10$, *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-

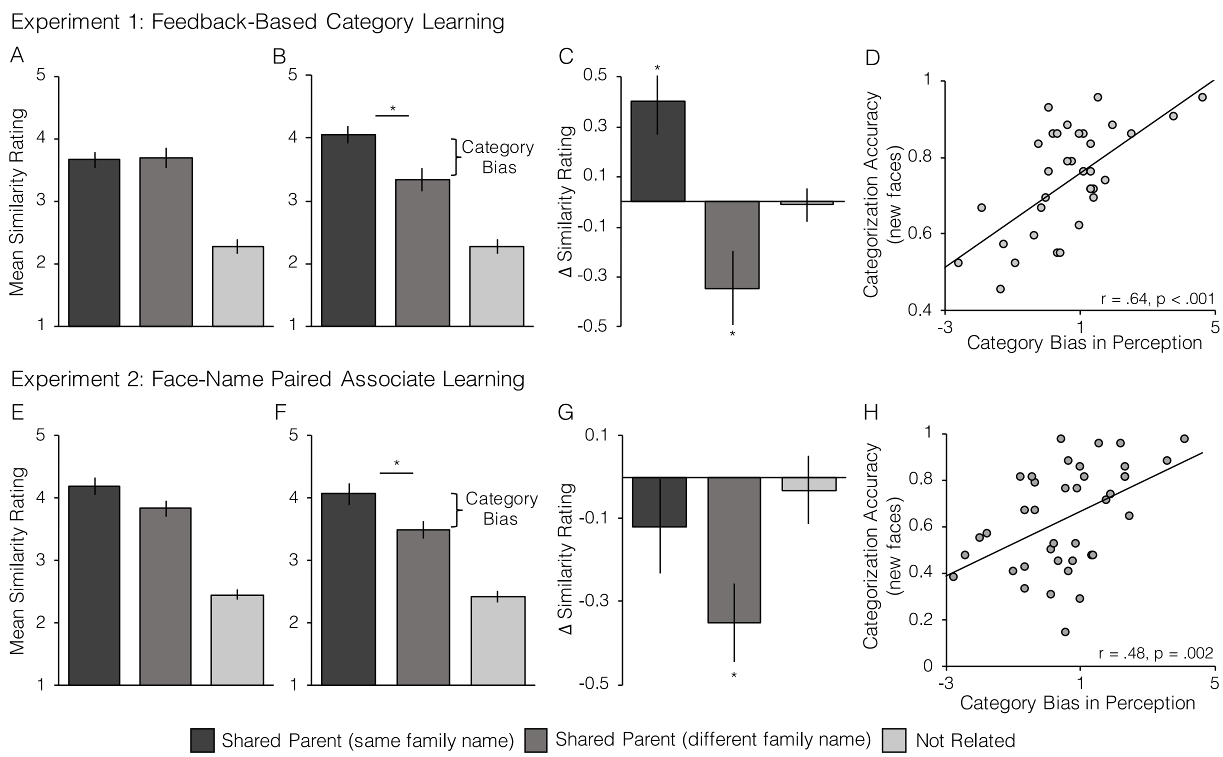

Figure 2. Top panel are results from the traditional category learning experiment. Bottom panel (shaded grey) are results from the face-name paired associate learning experiment. ***A & E.*** Average similarity ratings for faces that share a parent and family name, faces that only share a parent, and faces that don't share any parents before learning. ***B & 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 & G.*** Changes in similarity ratings from pre- to post-learning. Asterisk denotes significant ($p < .05$) increases and decreases in perceived similarity for faces. ***D & H.*** Positive relationship between indirect (category bias in perception) and direct (categorization accuracy for new faces) measures of memory generalization.

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 x 2 (timepoint [pre-learning, post-learning] x pair-type [shared parent-same family name, shared parent-different family name]) repeated-measures ANOVA showed only a marginal interaction between timepoint and condition ($F(1, 38) = 2.87$, $p = .098$, $\eta_p^2 = .07$). 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's 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 $\beta = .30$, $t(34) = 1.80$, $p = .08$; post-learning category bias $\beta = .46$, $t(34) = 2.75$, $p = .01$).

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 $\beta = -.22$, $t(38) = -0.86$, $p = .40$; post-learning category bias $\beta = .66$, $t(38) = 2.57$, $p = .01$).

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. We were interested in how well people generalize family names to new faces in the two tasks and to what degree category bias in perceived similarity ratings indicates the formation of category knowledge prior to explicit generalization demands.

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 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, Gitelman, Parrish, & Mesulam, 2003). 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 even within the same category.

Learning-related changes in perceived similarity ratings were observed in both experiments. In Experiment 1, consistent with prior studies (Beale & Keil, 1995; Goldstone, 1994a, 1994b; Goldstone et al., 2001; Rosch & Mervis, 1975), similarity ratings for faces within a family increased while similarity ratings for faces that were physically similar but belonged to different families decreased. These shifts in perceived similarity may reflect allocation of selective attention to features that are category-relevant while diverting attention away from category-irrelevant features (Goldstone & Steyvers, 2001; Kruschke, 1996; Nosofsky, 1991). In contrast, in Experiment 2 the face-name paired-associate learning was associated with an overall decrease in similarity ratings from pre- to post-encoding, driven primarily by decreased similarity for faces that were physically similar but belonged to different families. This decrease in similarity ratings could reflect learning-related differentiation of representations to minimize confusability and interference (Chanales, Oza, Favila, & Kuhl, 2017; Favila, Chanales, & Kuhl, 2016; Hulbert & Norman, 2015; Kim, Norman, & Turk-Browne, 2017; Lohnas et al., 2018). Changes in perceived similarity ratings were modulated by category membership of the faces in both experiments, indicating that people tended to link faces with a shared last name even outside the context of a traditional category learning task.

The inclusion of similarity ratings also allowed us to address the question of whether or not people spontaneously link related information in service of generalization prior to explicit generalization demands. The category bias in similarity ratings observed after learning predicted subsequent generalization of category information to new examples in both experiments, indicating that both measures index the same category knowledge formation. Critically, the category bias was measured *after* learning but *before* the explicit generalization test, indicating that people likely linked related faces at encoding (see also Shohamy & Wagner, 2008; Zeithamova, Dominick, & Preston, 2012) rather than in response to generalization demands. Our results also extend prior studies on changes in perceived similarity as a result of explicit instruction where attention is directed towards category-relevant similarities (Goldstone, 1994b, 1994a; Livingston et al., 1998) to a novel task where attention was directed towards individuating differences. Observation of the category bias after the face-name paired-associate learning indicates that the mere presence of a shared piece of information biased perceived similarity in many participants.

In summary, our findings indicate that generalizable category representations form at encoding, prior to explicit generalization demands. Individuals spontaneously linked related experiences to form conceptual knowledge even when learning goals required participants to learn individuating differences between stimuli. The relationship between category bias in similarity ratings and subsequent generalization further indicates that measures of perceived similarity are useful for measuring category learning without explicit demands to generalize. Building upon long lines of research on category learning (*for reviews see* Ashby & Maddox, 2011; Seger, 2008) and categorical perception (Etcoff & Magee, 1992; Liberman, Harris, Hoffman, & Griffith, 1957; Livingston et al., 1998; *for reviews see* Goldstone & Hendrickson, 2010; Harnad, 2006), the current work links generalization and perception together and extends prior findings beyond traditional category learning paradigms.

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/?view_only=ca5a189813b14dfebd9804151bc1a1ed).

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

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

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

Chanales, A. J. H., Oza, A., Favila, S. E., & Kuhl, B. A. (2017). Overlap among spatial memories triggers repulsion of hippocampal representations. *Current Biology*, *27*, 1–11. https://doi.org/10.1016/j.cub.2017.06.057

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, *44*(3), 227–240. https://doi.org/10.1016/0010-0277(92)90002-Y

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

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

Favila, S. E., Chanales, A. J. H., & Kuhl, B. A. (2016). Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nature Communications*, *7*, 11066. https://doi.org/10.1038/ncomms11066

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

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., 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

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

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. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101.

https://doi.org/10.3758/BF03202365

Hulbert, J. C., & Norman, K. A. (2015). Neural differentiation tracks improved recall of competing memories following interleaved study and retrieval practice. *Cerebral Cortex*, *25*(10), 3994–4008. https://doi.org/10.1093/cercor/bhu284

Kim, G., Norman, K. A., & Turk-Browne, N. B. (2017). Neural differentiation of incorrectly predicted memories. *The Journal of Neuroscience*, *37*(8), 2022–2031. https://doi.org/10.1523/JNEUROSCI.3272-16.2017

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. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

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

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

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368. https://doi.org/10.1037/h0044417

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

Lohnas, L. J., Thesen, T., Doyle, W. K., Devinsky, O., Duncan, K., & Davachi, L. (2018). Timeresolved neural reinstatement and pattern separation during memory decisions in human hippocampus. *Proceedings of the National Academy of Sciences*, *115*(31), E7418–E7427. https://doi.org/10.1073/pnas.1717088115

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

Nosofsky, R. M. (1988). Exemplar-Based Accounts of Relations Between Classification, Recognition, and Typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708. https://doi.org/10.1037/0278-7393.14.4.700

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.

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

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

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

Rosch, E., & Mervis, C. B. (1975). Family resemblances. *Cognitive Psychology*, *7*, 573–605.

https://doi.org/10.1186/gb-2002-3-12-reports0063

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B*, *372*(1711), 20160049. https://doi.org/10.1098/rstb.2016.0049

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

Schlichting, M. L., Zeithamova, D., & Preston, A. R. (2014). CA1 subfield contributions to memory integration and inference. *Hippocampus*, *24*(10), 1248–1260. https://doi.org/10.1002/hipo.22310

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

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

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*(2), 195–231. https://doi.org/10.1037/0033-295X.99.3.582

Teyler, T. J., & DiScenna, P. (1986). The hippocampal memory indexing theory. *Behavioral Neuroscience*, *100*(2), 147–154. https://doi.org/10.1037/0735-7044.100.2.147

Winocur, G., Moscovitch, M., & Sekeres, M. (2007). Memory consolidation or transformation: Context manipulation and hippocampal representations of memory. *Nature Neuroscience*, *10*(5), 555–557. https://doi.org/10.1038/nn1880

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

Zeithamova, D., Schlichting, M. L., & Preston, A. R. (2012). The hippocampus and inferential reasoning: Building memories to navigate future decisions. *Frontiers in Human Neuroscience*, *6*, 1–14. https://doi.org/10.3389/fnhum.2012.00070