Improving Category Learning through Graded Classification

Mercury Mason (mmason2@binghamton.edu)
Department of Psychology, 4400 Vestal Parkway East,
Binghamton, NY 13902-6000 USA

Kenneth J. Kurtz (kjk@binghamton.edu)
Department of Psychology, 4400 Vestal Parkway East,
Binghamton, NY 13902-6000 USA

Abstract

Real-world categories often exhibit graded structure, yet learners struggle to acquire family resemblance categories compared to unidimensional ones in laboratory studies. We propose that part of this difficulty arises from the binary nature of Traditional Artificial Classification Learning. We introduce Graded Classification Learning, a paradigm integrating category and quality judgments into response and feedback phases of a learning trial. This allows higher fidelity feature space exploration, aligning more with naturalistic learning processes. The 'graded' learners showed superior performance, with higher final accuracy and steeper learning curves than the 'traditional' learners. While aggregate response patterns appeared similar across conditions, profile analysis revealed an apparent gradedness in the traditional condition that was masked by an overwhelming preference for (bisecting) unidimensional strategies, whereas graded participants mostly exhibited genuine graded responses. These findings suggest traditional binary tasks may inadvertently hinder learning of graded structure and that incorporating quality judgments fosters robust category representations.

Keywords: concepts and categories; category learning; classification; knowledge representation

Introduction

Categorization is a core cognitive process that accounts for our ability to encode, represent, and make inferences about the world in a structured way. Categories in the natural world rarely conform to simple rules based on a single feature. Instead, they typically exhibit graded structure characterized by family resemblance relationships, where category members share overlapping features without any single attribute being necessary or sufficient for membership (Rosch & Mervis, 1975; Wittgenstein, 1953). This structure reflects the probabilistic nature of real-world concepts, where feature correlations and frequency distributions often shape category boundaries, with quality of membership within a category being a function of the degree of featural similarity with other category members. Early evidence indicated humans form graded representations, showing effects driven by typicality and quality statistics within such category structures (Rosch, 1973b; Rosch et al., 1976).

A straightforward hypothesis would follow from the previous summary: when confronted with family resemblance structure, humans are well equipped to place equal weight on all dimensions when learning or forming categories that reflect said structure. However, a considerable body of evidence suggests dimensional biases frequently influence categorization behavior, often at the expense of category coherence. The bias toward unidimensional (UNI) solutions when family resemblance is present is especially pronounced in unsupervised sorting tasks where there is no error-driven feedback (Lassaline & Murphy, 1996; Medin et al., 1987; Milton & Wills, 2004, 2009; Patterson et al., 2019; Regehr & Brooks, 1995; Spalding & Murphy, 1996). In supervised categorization tasks (e.g., classification with feedback) acquire simpler rule-based categories faster than family resemblance categories (Kurtz et al., 2013; Nosofsky et al., 1994; Shepard et al., 1961).

We consider a driver of simpler, yet less satisfactory unidimensional strategy adoption for family resemblance categories. We propose that part of this difficulty arises from mismatches between naturalistic category learning and the constraints of traditional laboratory tasks. Specifically, the binary (correct/incorrect) feedback in traditional artificial classification learning (TACL) tasks may fail to engage the distributed feature comparison processes critical for detecting family resemblance structure. While laboratory tasks necessarily simplify real-world complexity, this dichotomous framing contrasts with aspects of naturalistic categorization where individuals often evaluate not just whether an item is a category member, but also how 'good' or typical an example it is (Rosch, 1975), a judgment crucial for guiding inferences beyond simple labeling. Computational models like SUSTAIN (Love et al., 2004) show that tasks requiring only category labels encourage learners to form deterministic rules rather than probabilistic feature clusters. Similarly, auxiliary goals to classification, like feature inference and typicality ratings, appear to better promote access to feature correlations (Ahn et al., 2002; Chin-Parker & Ross, 2004). These observations align with a broader literature demonstrating that the specific demands of the learning task (e.g., classification vs. inference, observation vs. active responding) can profoundly shape what is learned about category structure (Markman & Ross, 2003; Yamauchi & Markman, 1998). Even supervised observation (no response required) yields greater sensitivity to the distributional properties of features than classification (Levering & Kurtz, 2015).

In typical TACL tasks, participants receive a visual stimulus, make binary category judgments (e.g., "A" or "B")

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

and receive binary feedback (correct/incorrect). This dichotomy contrasts sharply with natural categorization, where items are often evaluated as better/worse exemplars based on feature distributions and typicality gradients (Rosch, 1973a; Murphy, 2004). Crucially, the constrained response and feedback structure of TACL, where only categorical correctness is emphasized, drives learners toward strategies which allow them to effectively discriminate between members of one category and another (Kurtz, 2015).

While discriminative learning is useful for predicting labels and drawing decision boundaries, it fails to model the distributional properties that underlie each class within a domain – arguably, a core element of human categorization. Generative learning, which emphasizes this kind of model building, speaks to the goal of understanding the internal structure of a category (Ng & Jordan, 2001). Though TACL is not entirely discriminative, learning modes that foster generative knowledge are worth exploring for both theoretical and applied ends.

To address these limitations, we introduce graded classification learning (GCL), a novel paradigm designed to re-engage the feature integration processes disrupted by traditional tasks. In GCL, participants provide responses that incorporate both category membership and graded quality judgments (e.g., "Very Good A," "Acceptable B") while receiving feedback on both components. Learners must predict the class label and determine its position within the topology of a category's manifold. By incorporating quality judgments, GCL prompts learners to evaluate feature interactions by assessing how well a given exemplar aligns with the internal structure of each category. In this way, GCL aims to incorporate elements more aligned with how people naturally learn and use categories in the real world, particularly the evaluation of category member 'goodness' or typicality. When encountering new instances of a category, people rarely make purely binary choices. Instead, they evaluate how well items fulfill their roles as category members. For example, when learning about chairs, people consider not just identification, but how well it functions, its typicality, and relation to prior experiences. These 'secondary' evaluations are fundamental to building robust conceptual models. With GCL, we strive to better support this kind of processing natively in the task and better capture the natural tendency to not only encode an item based on its visible properties, but also encode it with respect to its brother and sister members.

We hypothesize that GCL will foster robust representations of graded structure, improving both learning and generalization to novel exemplars. The latter serves as a critical marker of ecological validity, reflecting naturalistic category use where feature correlations guide typicality and inference (Chin-Parker & Ross, 2002; Rehder & Burnett, 2005). By prompting learners to consider not just 'which category?' but also 'how good an example?', GCL may engage deeper representational processes sensitive to within-category structure, a hallmark of how humans often use categories (Medin & Schaffer, 1978). To evaluate this hypothesis, we address three interrelated questions. First, we test whether GCL enhances family resemblance category acquisition compared to TACL. Second, we assess whether GCL and TACL produce distinct representational profiles as they related to the underlying category structure. Third, we examine how training methods affect downstream flexibility in tasks requiring graded judgments (e.g., typicality ratings) and feature inference, processes central to real-world categorization. By unifying these lines of inquiry, we evaluate the broader claim that paradigm-induced constraints greatly influence the behavioral predisposition to acquire FR categories.

Method

Participants

A total of 111 undergraduate students from Binghamton University participated for partial credit toward a course requirement. Participants were randomly assigned to either the traditional (TRAD) (n=55) or graded (GRADED) (n=56) condition.

Materials and Design

Stimuli consisted of hexagonal "computer chips" generated within a continuous two-dimensional feature space defined by brightness (1–9 scale: 1 = darkest (rgb[28,28,97]), 9 = lightest (rgb[197,198,246]) and size (1–9 scale: 1 = largest, 9 = smallest). Each stimulus took on a purple hue with a black outline and was presented against a white background. These dimensions formed a 9×9 grid where coordinates (1, 1) represented the darkest/largest hexagon and (9, 9) the lightest/smallest. This continuous two-dimensional implementation of family resemblance structure offers several methodological advantages over traditional discrete-feature approaches. First, the continuous feature space enables precise measurement of how learners weight and integrate multiple dimensions during category formation, as opposed to binary feature presence/absence paradigms common in previous work. Second, the systematic sampling of the feature space allows for controlled investigation of category boundaries and typicality gradients, critical for understanding how learners partition the space. Third, the geometric relationship between features creates natural opportunities to test generalization through interpolation (novel combinations of trained feature values) and extrapolation (extension to more extreme feature values).

Two FR categories, arbitrarily labeled "Tesla" and "Fermi", were constructed around anchor points for training: Tesla (generally dark/large) at (2, 2) and Fermi (generally light/small) at (8, 8). From the 81 total feature combinations, 30 exemplars (15 per category) were selected from non-overlapping regions of the feature space. An additional 35 exemplars (interpolation, extrapolation, and boundary items) were selected and held out during training to assess generalization during testing. Category quality levels for the GRADED condition were determined by Manhattan distance from ideals: 'Very Good' members included items at or one

Manhattan block from the ideal (e.g., Tesla: (2, 2), (2, 3)), 'Good' members spanned 2–3 blocks (e.g., Tesla: (4, 2), (2, 5)), and 'Acceptable' members occupied 4–5 blocks (e.g., Tesla: (6, 2), (2, 7)).

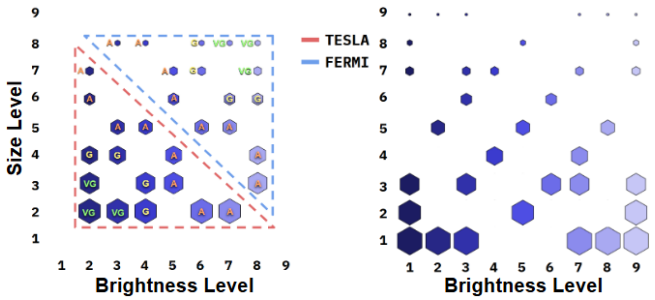

Figure 1: Training items for 'Tesla' (red region) and 'Fermi' (blue region) (Left). Novel generalization items (extrapolation, interpolation, boundary) shown to participants during testing (Right). Note: Stimuli presented here are not to scale. The figure is meant to illustrate the distribution of the feature space.

This study employed a between-subjects design comparing two training conditions. In the TRAD condition, participants made binary category judgments ("Tesla" or "Fermi") using a two-level slider and received feedback limited to categorical correctness (correct/incorrect). In the GRADED condition, participants used a six-level slider combining category membership and quality judgments ("Very Good," "Good," or "Acceptable" × "Tesla/Fermi") and received feedback on both category accuracy and quality appropriateness (see Table 1). For analyzing learning curves and final accuracy, responses were scored for class correctness ('Tesla'/'Fermi'). Thus, chance performance for this accuracy measure was 50% in both conditions. (For GRADED, chance on the full combined category-quality judgment was 16.7%, but class correctness was the primary measure for learning curves).

Procedure

The experiment comprised three phases conducted sequentially. During the *training phase*, participants completed six blocks of 30 trials each (15 trials per category), with stimuli presented in randomized order. Each trial began with a central fixation cross (500 ms), followed by stimulus presentation until a condition-specific classification response was made using the slider. Condition-specific feedback text appeared immediately following the response and was available until the participant clicked a "next" button, which began the next trial.

Following training, participants entered the *test/generalization phase* comprising 65 trials (30 trained items, 35 novel items). The novel items were strategically selected to probe different aspects of category learning through three types of generalization challenges. Interpolation items tested participants' ability to categorize new combinations of previously encountered feature values, assessing whether they had developed integrated representations of the feature space. Extrapolation items, involving more extreme feature values, evaluated how participants extended their category knowledge beyond the trained regions. Boundary items, positioned between category prototypes, were particularly diagnostic of learning outcomes: participants who acquired genuine family resemblance structure should show graded typicality responses reflecting distance from category prototypes, while those using simple classification rules should display sharp category boundaries. For each trial, participants first classified the item (Tesla/Fermi) via binary classification[1] and then provided a typicality rating on a 7-point scale (1 = "Not Very Typical" to 7 = "Very Typical") for their chosen class. No feedback was provided during this phase.

The final *feature inference phase* required participants to infer missing feature values across 18 trials. On each trial, a randomly selected category label and feature dimension (brightness or size) were displayed. Participants selected the missing feature from four random equidistant options spanning the 9-level dimension that would combine with the given feature to create a member of the given category. For example, given "Tesla" and brightness = 2, size options might include levels 3, 5, 8, and 9. Responses were self-paced, with no feedback provided.

Table 1. Examples of Feedback by Condition

Condition	Response Status	Feedback
TRAD	Correct	*CORRECT! You answered "Tesla". This is a member of the Tesla Category.*
	Incorrect	*INCORRECT! You answered "Fermi". This is a member of the Tesla Category.*
GRADED	Correct class, correct quality	*CORRECT! You answered "Good Tesla". This is a member of the Tesla category and it is a Good Tesla.*
	Correct class, incorrect quality	*CORRECT! You answered "Very Good Tesla". This is a member of the Tesla category. However, it is better considered as a Good Tesla.*
	Incorrect class, correct quality	*INCORRECT! You answered "Good Fermi". This is a member of the Tesla category. In fact, it is best considered as a Good Tesla.*
	Incorrect class, incorrect quality	*INCORRECT! You answered "Acceptable Fermi". This is a member of the Tesla category. In fact, it is best considered as a Good Tesla.*

[1] We acknowledge that binary classification testing could advantage TRAD participants through transfer-appropriate processing (Franks et al., 2000). However, this test design (binary classification followed by typicality rating for both groups) enables direct performance comparisons. An alternative, such as having GRADED participants give graded responses at test, would have likely impacted the subsequent typicality ratings in a confounding manner.

Results

Analysis of learning curves (accuracy on class responses only) revealed marked differences in category acquisition between conditions. A mixed ANOVA with condition (GRADED vs. TRAD) as a between-subjects factor and block (1–6) as a within-subjects factor demonstrated a significant interaction effect, $F(5, 545) = 3.42$, $p = .01$, $\eta^2 = 0.03$. This interaction indicates that the rate of learning across blocks differed significantly between conditions; GRADED participants exhibited a steeper learning trajectory, culminating in significantly higher accuracy by the final training block (91% vs 83%), t(109) = 2.03, p = .048, Bonferroni-corrected. While this p-value is close to the .05 threshold, the finding is statistically significant after correction for multiple comparisons, suggesting accelerated category acquisition under the enriched response-feedback conditions.

Aggregate choice behavior during testing appeared strikingly similar (see figure 2, Top) across conditions. At first glance, this would seem to suggest a high degree of representational similarity produced by both tasks. However, these aggregate measures mask fundamental qualitative differences in what representations and in what proportions these representations existed in our sample. While both conditions showed overall improvement across training blocks, examining individual response profiles revealed that similar performance levels overall emerged from markedly different category representations. This highlights a critical limitation of relying solely on group-level accuracy metrics in category learning research: superficially similar learning curves can arise from fundamentally different cognitive representations.

To characterize these underlying representational strategies, response patterns were visually classified into three profiles: *FR* (integration of brightness and size), *UNI-X* (brightness-dominant), and *UNI-Y* (size-dominant). Chi-square analysis revealed a significant condition-strategy association ($\chi^2(2) = 24.31$, $p < .001$). GRADED participants predominantly adopted FR strategies (69%, $n = 33$), with fewer unidimensional users (UNI-X: 15%, UNI-Y: 17%). In contrast, TRAD participants overwhelmingly relied on unidimensional rules (UNI-X: 38%, UNI-Y: 45%), with only 17% ($n = 8$) demonstrating FR profiles. Individual profiles that did not clearly fit any of the three theoretically relevant strategies (9 participants in TRAD, 8 participants in GRADED) were not included in this specific profile-based chi-square analysis. These excluded profiles generally showed inconsistent or noisy response patterns that did not align clearly with unidimensional or fully graded strategies.

Typicality ratings provided critical insight into implicit category structure. While unidimensional strategists in both conditions prioritized their dominant dimension during classification, their typicality ratings retained sensitivity to the secondary dimension (e.g., UNI-X TRAD participants rated larger Tesla exemplars as more typical [$r = 0.52$, $p < .001$]). This dissociation suggests that unidimensional strategies reflect task-driven response simplification rather than completely impoverished representations.

Feature inference performance further distinguished conditions. GRADED participants, particularly true graded responders, selected missing features more aligned with category prototypes (mean deviation from ideal: 1.2 vs. 2.7 levels in TRAD; $t(109) = 4.11$, $p < .001$) and exhibited stronger sensitivity to feature correlations (GRADED: $r = 0.68$ vs. TRAD: $r = 0.39$, $p = .004$). Notably, 22% of TRAD responses invalid responses (e.g., selecting brightness = 9 for Tesla), compared to 6% in GRADED ($\chi^2(1) = 8.92$, $p = .003$), further demonstrating a GRADED advantage.

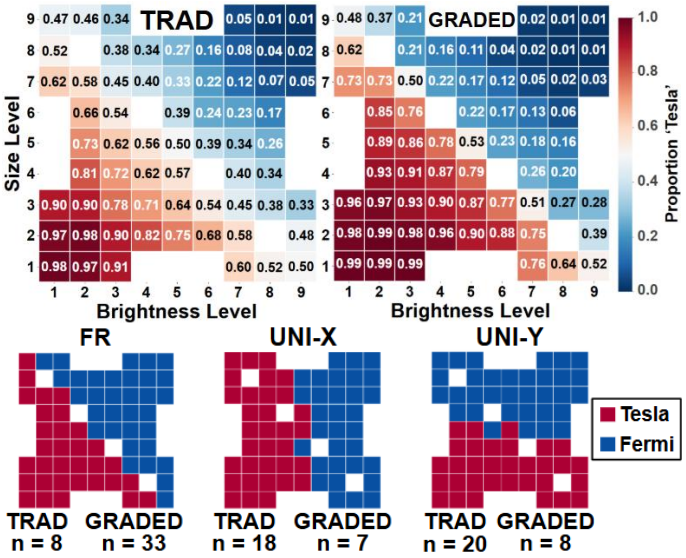

Figure 2: (Top) Aggregate Test Heatmaps by Condition. TRAD aggregate test heatmap suggests participants learned graded representations, but individual profiles reveal the aggregate masked a high proportion of UNI learners. GRADED aggregate test heatmap suggests participants learned distributed representations along both dimensions which was supported by a majority of FR individual profiles. (Bottom) Test Heatmap Profile Type by Condition. UNI profiles (UNI-X, UNI-Y) were much more common in the TRAD condition, while FR profiles made up a strong majority in the GRADED condition. Note: Displayed profiles serve as example 'templates' to illustrate profile types.

Discussion

The challenge of acquiring family resemblance (FR) categories in laboratory settings, despite their ecological prevalence, has long suggested a disconnect between natural and experimental learning environments. Our findings demonstrate that this disparity arises not from inherent processing limitations, but from mismatches between traditional task designs and the graded nature of real-world categorization. By introducing Graded Classification Learning (GCL), which incorporates quality judgments and multidimensional feedback into a classification paradigm, we observed marked improvements in category acquisition, representational fidelity, and representational robustness. These results support the intuition that Traditional Artificial Classification Learning (TACL), while a staple of the study

of human category learning, encourages more discriminative learning strategies that might not be optimal for all feature spaces (Kurtz, 2015).

The stark strategy adoption contrast indicates response-feedback structure profoundly influenced category representation development. Unidimensional profiles pervaded TRAD (83%), while graded profiles were common in GRADED (69%). This echoes findings that task demands shape acquisition and system activation (Ashby & Maddox, 2005; Seger & Miller, 2010). Binary feedback encouraged simplified rule-based strategies; graded feedback promoted sophisticated, similarity-based representations better capturing category dimensionality. This plasticity highlights adaptive human categorization: learners optimize strategies to task structure, often at the expense of ecological validity in traditional paradigms. Particularly revealing is the dissociation between classification behavior and typicality ratings among TRAD participants. Though often defaulting to unidimensional classification, their typicality judgment sensitivity to secondary features suggests richer mental representations than their classification indicated. This implies unidimensional responding was a strategic adaptation to task demands, not a fundamental representational limit; TRAD participants may have optimized for classification verbalization, while continuous typicality ratings better fit their underlying representations. GCL, by making membership gradations task-relevant, appears to bridge this implicit-explicit knowledge gap. Feature inference results provide perhaps the strongest evidence for qualitative representational differences. GRADED participants' superior missing feature inference and stronger feature correlation sensitivity suggest more cohesive, prototypical representations. GRADED's lower invalid feature selection rate further indicates graded feedback promotes more stable, well-defined category boundaries.

Our methodology proved valuable. The continuous structure allows for precise tracking of feature weighting/use, revealing fine-grained strategic differences obscured by binary-feature designs. This highlights stimulus space design's importance: continuous spaces better capture how learners partition feature dimensions and establish boundaries (valuable for FR acquisition) than discrete features identifying only broad patterns. Critically, profile analysis revealed that aggregate performance masked different underlying representational strategies. This group-individual dissociation underscores a key limitation of summary metrics: obscuring distinct representational geometries (Estes, 1956; Siegler, 1987; Smith & Minda, 2000). Consequently, this finding invites both a theoretical reexamination of previous category learning studies that relied primarily on aggregate measures and suggests new directions for future work that incorporate detailed individual analyses to uncover the impact of learning contexts on representation.

Our 2D stimulus space offered control, but natural categories are higher-dimensional. Future work should test if GCL's benefits scale with dimensionality or interact with perceptual salience (reducing selective attention) Disentangling graded response vs. graded feedback contributions is also key: is multidimensional prediction or feedback driving the effect? Integrating GCL with unsupervised paradigms (Love, 2002; Pothos et al., 2011) or exploring continuous GCL versions could clarify if information gain or exploration strategies are crucial. Continuous versions of GCL may also be worth exploring, especially if what drives a GCL advantage is information gain or exploration strategies.

These findings may prompt re-evaluation of theoretical models and could inform educational practices. Theoretically, they question classic associative accounts that reduce category learning to statistical processing, demonstrating instead that response and feedback format shapes not just learning rate but also the fundamental structure of mental representations. While binary classification is not a primary classroom teaching method, GCL's principle of incorporating graded evaluations offers insights for educational settings desiring nuanced understanding over binary correctness (Anderson et al., 1995); for instance, prompting students to evaluate an example's 'strength' could foster deeper conceptual understanding. Overall, this work compellingly shows that integrating graded responses and feedback, rather than binary choices, encourages greater learning and representation of family resemblance. Crucially, this study provides a platform for subsequent research into learning modes beyond traditional paradigms, toward tasks promoting more generative knowledge acquisition.

References

Ahn, W. K., Marsh, J. K., Luhmann, C. C., & Lee, K. (2002). Effect of theory-based feature correlations on typicality judgments. *Memory & Cognition*, *30*, 107-118.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, *4*(2), 167-207.

Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Memory & Cognition*, *30*(1), 119-128.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol.*, *56*(1), 149-178.

Chin-Parker, S., & Birdwhistell, J. (2017). Category learning with a goal: how goals constrain conceptual acquisition. *Journal of Cognitive Psychology*, *29*(4), 450-468.

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: a comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 216.

Chin-Parker, S., & Ross, B. H. (2002). The effect of category learning on sensitivity to within-category correlations. Memory & cognition, 30(3), 353-362.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134.

Franks, J. J., Bilbrey, C. W., Lien, K. G., & McNamara, T. P. (2000). Transfer-appropriate processing (TAP). *Memory & Cognition*, *28*, 1140-1151.

Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. In *Psychology of learning and motivation* (Vol. 63, pp. 77-114). Academic Press.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 552.

Lassaline, M. E., & Murphy, G. L. (1996). Induction and category coherence. *Psychonomic Bulletin & Review*, *3*(1), 95-99.

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition*, *43*, 266-282.

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

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological Review*, *111*(2), 309.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological bulletin*, *129*(4), 592.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207-238.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*(2), 242-279.

Milton, F., & Wills, A. J. (2009). Long-term persistence of sort strategy in free classification. *Acta Psychologica*, *130*(2), 161-167.

Milton, F., & Wills, A. J. (2004). The influence of stimulus properties on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 407.

Minda, J. P., & Ross, B. H. (2004). Learning categories by making predictions: An investigation of indirect category learning. *Memory & cognition*, *32*(8), 1355-1368.

Murphy, G. (2004). *The big book of concepts*. MIT press.

Ng, A., & Jordan, M.I. (2001). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. *Neural Information Processing Systems*.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*(3), 352-369.

Patterson, J. D., Snoddy, S., & Kurtz, K. J. (2019). Family Resemblance in Unsupervised Categorization: A Dissociation Between Production and Evaluation. *Cognitive Science* (pp. 2537-2543).

Pothos, E. M., Edwards, D. J., & Perlman, A. (2011). Supervised versus unsupervised categorization: Two sides of the same coin?. *Quarterly Journal of Experimental Psychology*, *64*(9), 1692-1713.

Regehr, G., & Brooks, L. R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 347.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive psychology*, *50*(3), 264-314.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, *104*(3), 192-233.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328-350.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. *Cognitive development and the acquisition of language/New York: Academic Press*.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573-605.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 491.

Seger, C. A., & Miller, E. K. (2010). Category learning in the brain. *Annual review of neuroscience*, *33*(1), 203-219.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of experimental psychology: General*, *116*(3), 250-264.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411.

Smith, D. J., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 3.

Spalding, T. L., & Murphy, G. L. (1996). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(2), 525.

Wittgenstein, L. (1953). Philosophical Investigations. Philosophische Untersuchungen. Oxford, England: Macmillan.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, *39*(1), 124-148.