Switch it up: Learning Categories via Feature Switching

Garrett Honke (ghonke1@binghamton.edu)
Nolan Conaway (nconawa1@binghamton.edu)
Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, Binghamton University (SUNY)
Binghamton, NY 13905 USA

Abstract

This research introduces the switch task, a novel learning mode that fits with calls for a broader explanatory account of human category learning (Kurtz, 2015; Markman & Ross, 2003; Murphy, 2002). Learning with the switch task is a process of turning each presented exemplar into a member of another designated category. This paper presents the switch task to further explore the contingencies between learning goals, learning modes, outcomes, and category representations. The process of successfully transforming exemplars into members of a target category requires generative knowledge such as within-category feature correspondences – similar to inference learning. Given that the ability to switch items between categories nicely encapsulates category knowledge, how does this relate to more familiar tasks like inferring features and classifying exemplars? To address this question we present an empirical investigation of this new task, side-by-side with the well-established alternative of classification learning. The results show that the category knowledge acquired through switch learning shares similarities with inference learning and provides insight into the processes at work. The implications of this research, particularly the distinctions between this learning mode and well-known alternatives, are discussed.

Keywords: concepts; learning; categorization; category use

Introduction

In light of recent work establishing the effects of category use on conceptual representation, it has been argued that studying human category learning solely through the traditional artificial classification learning (TACL) paradigm provides a limited view of the processes involved (Chin-Parker & Ross, 2002; Levering & Kurtz, 2015; Love, 2002; Yamauchi & Markman, 1998; see Kurtz, 2015 for review). Specifically, category representations acquired through TACL are in many ways specially tailored for performing the classification task itself—a procedure that requires discriminating between a set of different categories based on the features of individual examples. Considering that real-world concepts commonly serve functions beyond class discrimination, there is clear importance for developing theories of human category learning that generalize across different types of learning goals and opportunities (Murphy, 2002).

In the spirit of evaluating how distinct category representations can arise under alternative learning conditions, we introduce the *switch task* – a new category learning technique based on transforming exemplars into members of another category – inspired by the DIVA account of human category learning (Kurtz, 2007, 2015). The chief goals of this paper are to introduce the switch task and present empirical data exploring the differences in category representation that arise as a result of this learning mode.

Research Motivation: The DIVA Model

DIVA (Kurtz, 2007) is an artificial neural network (ANN) model that uses a DIVergent Autoencoder architecture trained via backpropagation (Rumelhart, Hinton, & Williams, 1986) to learn and represent categories. The model is a concrete instantiation of a theory for how humans learn and represent categories – namely, that psychological categories are task-constrained, generative models of the regularities that exist among a category's members (Kurtz, 2015). One novel property of the model is that, within the context of a classification problem, category representations are not built independently (see Figure 1).

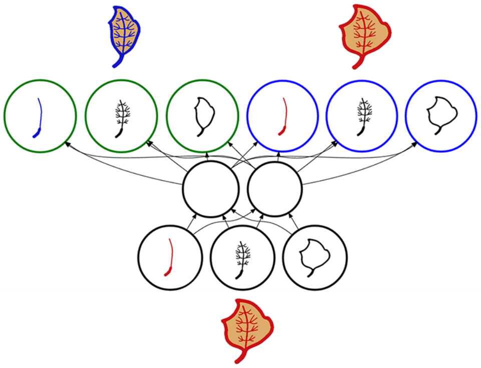


Figure 1: The DIVA model architecture with two possible reconstructions of a category exemplar. Stimuli are drawn from the present study.

The switch task is motivated by this unique design feature where each stimulus is reconstructed as a possible member of each category in the task. At the core of switch learning is an invitation to construe each presented example as a member of an alternate category, and then transform the example accordingly by changing its feature values as required. Although DIVA does not actually produce switch responses or learn categories by switching features, the key motivation of this work is to determine whether and with what properties the process of generating alternate categorical construals promotes category acquisition. To properly characterize how learning categories with the switch paradigm might create distinct category representations, the next section addresses the common properties of category knowledge after learning via classification.

The TACL paradigm

In the dominant TACL approach to studying human category learning, observers are presented with exemplars belonging to a set of experimenter-defined categories. For each observation, learners are asked to make a classification decision and are then provided corrective feedback. This guess-and-correct process is continued for a set number of iterations (or until a criterion is met) and then participants are tested on what they have learned.

An observed result of learning categories under this approach is that the task biases the resultant structure of conceptual representation (Solomon, Medin, & Lynch, 1999). In particular, individuals completing the classification task learn to select a category label based on an example with a particular set of features from among a set of alternatives. This has the effect of focusing the learner on features that have distinct values across the present categories (Chin-Parker & Ross, 2002). As such, learners need only to determine the features that are diagnostic for distinguishing between categories (when available).

Even small changes to the TACL paradigm can affect what is learned in the task. Levering and Kurtz (2015) have shown that removing the guess-and-correct component of the TACL paradigm increases the learning of within-category correlations — even when they are not useful towards predicting class membership. Likewise, asking learners to provide missing feature values (as opposed to class labels) promotes knowledge of within-category central tendencies (Markman & Ross, 2003; Yamauchi & Markman, 1998).

While the TACL approach has been favored for many valid reasons (notably, to reduce the scope of the problem and develop clear research questions), it also produces a fragmented view of the processes that underlie human category learning (Murphy, 2002, 2005; see Kurtz, 2015 for an in-depth exploration of these issues). As such, this project aims to widen current understanding of category learning task effects by presenting a comparison of two techniques: classification learning and switch learning.

The Current Study

In the present work, we primarily seek to evaluate the differences in conceptual representation fostered by the novel switch paradigm (fully described below) and traditional artificial classification learning. Specifically, questions of interest include whether either mode is better suited for different learning problems and different applications of category knowledge (i.e., inferring unobserved features, classifying partial items with features omitted).

We address these questions with an experiment consisting of a learning phase (switch or classification) and three test phases – switch, classification and inference tests. Participants learned about two categories of plant life, named *Lape* and *Tannet*, each consisting of four examples varying in three binary dimensions. The test phases included the eight examples from the learning phase plus partial exemplars. The classifications were three intermediate-difficulty category structures (Types II, III and IV; see Figure 2) from the elemental six-types problems (Shepard, Hovland, & Jenkins, 1961).

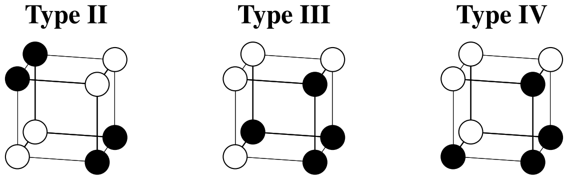


Figure 2: An instantiation of the experiment's category structures. The vertices represent individual exemplars with category marked by label and color; the spatial dimensions correspond to features of the stimuli.

The key difference between the switch task and other commonly studied learning modes is that participants are tasked with: 1) deciding which features of the example do not accord with the targeted category and 2) switching the binary values of these features so as to change the category membership. Participants use feature buttons to transform an initial example into a member of the target category. Each button shows an image of a feature value that is not currently present in the example. Examples are immediately updated on button press and the selected button is removed from the interface (the feature switch cannot be reversed). The only constraint on the switching task is that at least one feature must be switched to complete a trial. Accuracy feedback is provided at the end of each trial. The switch is considered accurate if the newly constructed example is a member of the target category.

Switch and classification training clearly differ in many respects – the central goal of this study is to evaluate the differences in conceptual representation conferred by two learning modes, as well as how these representations serve learners in putting their knowledge to work. If the process of viewing non-member exemplars as possible members of a target category helps people learn about underlying category structure, then the switch learning mode has the potential to be in some ways more effective than classification. Higher accuracy on test phases after learning would be clear evidence consistent with this hypothesis.

Alternatively, it is plausible that participants will be best at the test phase that mirrors their assigned learning condition. In accord with the construct of transfer-appropriate processing (Morris, Bransford, & Franks, 1977), this would suggest that classification learners will perform best on the classification test phase. This result would further validate a key idea guiding this research – that task conditions during category learning have a strong effect on category knowledge. Likewise, under this view, performance on the inference test phase has the highest importance: unlike the switch and classification tests, learners in both conditions have no experience completing inference trials in the context of our experiment.

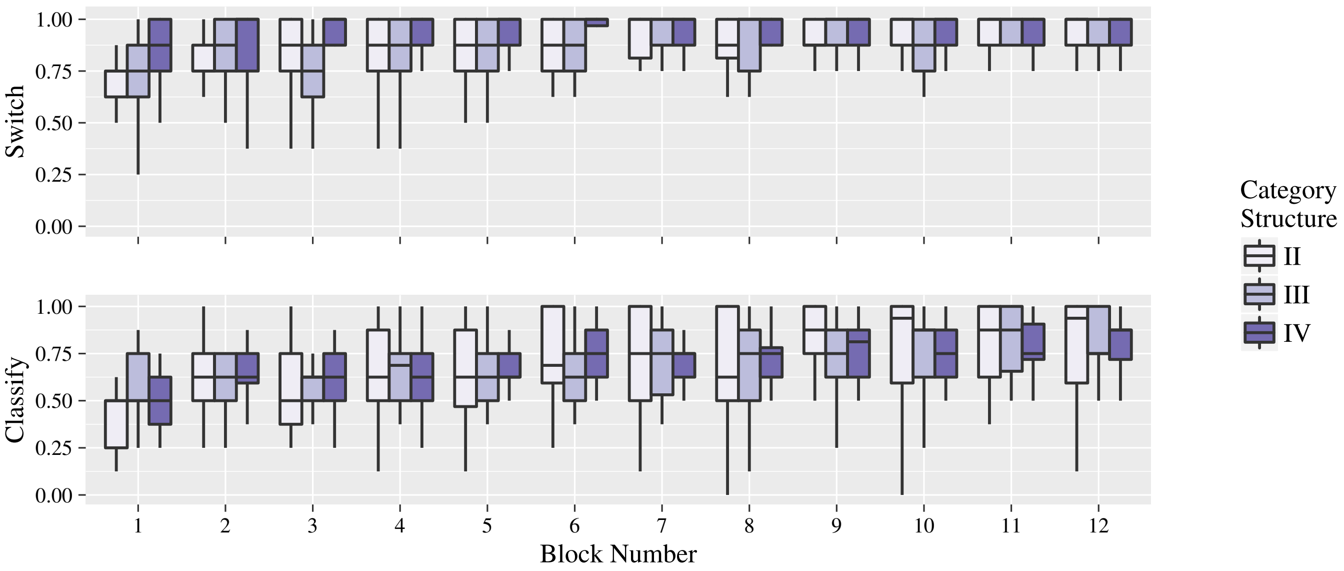


Figure 3: Mean percent correct for the training phase of the experiment. Switch learners were reliably better with Types III and IV as compared to Type II. No reliable differences were found within the classify group.

Method

Participants. All participants ($N = 170$) were recruited from the Binghamton University Psychology Department pool and randomly assigned to condition. Each participant gave consent to participate in writing and received credit towards the completion of a course requirement. Three participants were excluded from the analysis; two for failure to complete the experiment and one due to experimenter error.

Design and Procedure. There were two independent variables in the present study: learning condition and category structure. Three category types drawn from the Shepard et al. (1961) elemental category structures were included as a between-subjects variable. Learning condition was also a between-subjects variable – leading to a 2 (learning condition: classify, switch) x 3 (category structure: Type II, Type III, Type IV) design. The stimuli were leaf-like images (e.g., Figure 1) varying on three binary dimensions (color, veining, and shape). The assignment between conceptual and perceptual dimensions was counterbalanced across participants.

The experiment was conducted in private testing rooms on PCs with the use of a mouse and keyboard. The PsychoPy package was used for the development of the task interface (Peirce, 2007). Each participant was presented with instructions on screen. In the learning phase, participants completed 12 blocks of either classification or switch trials (8 trials per block) with feedback provided after every trial. Incorrect trials were repeated until a correct switch or classification was produced.

The goal for participants in the switch learning condition ($n = 84$) was to switch the features of provided examples until they matched a target category. The location of each feature button randomly varied by trial. After completing the transformation, participants used a Done button and then received feedback (see supplementary information[1] for a depiction of the switch interface).

Participants in the classification learning condition ($n = 86$) performed a task similar to that of the TACL paradigm except for the noted difference that incorrect trials were repeated until the correct category label was selected. On each trial, an example was presented and participants were asked to select the correct category label. After the classification decision was made, participants were given corrective feedback.

Three distinct test phases were included in the experiment: inference, classification and switch tests. All participants completed the inference test phase first and then the classification and switch phases (order randomly determined). On inference trials, learners were presented an incomplete example (one or two features omitted), a category label, and images of the two possible instantiations of one of the missing features as clickable buttons. Switch and classification test trials were identical to the training task. No feedback was provided during the test phases. The number of test trials in each phase depended on the category structure condition and test phase (see the supplementary materials for a depiction of the complete set of training and test examples, the task interface and all instructions[1]).

Results

The data was analyzed using generalized linear mixed effects regression (GLMER; Bates, Mächler, Bolker, & Walker,

[1] `gist.github.com/ghonk/7e24c78a05280f61e866`

2015) fit by maximum likelihood in the R analysis environment (R Core Team, 2016). The general analysis approach was to build regression models that predict trial success with learning condition (switch, classify), example type (complete training examples, incomplete novel examples) and their interaction. Study participants were included in the models as a random intercept.

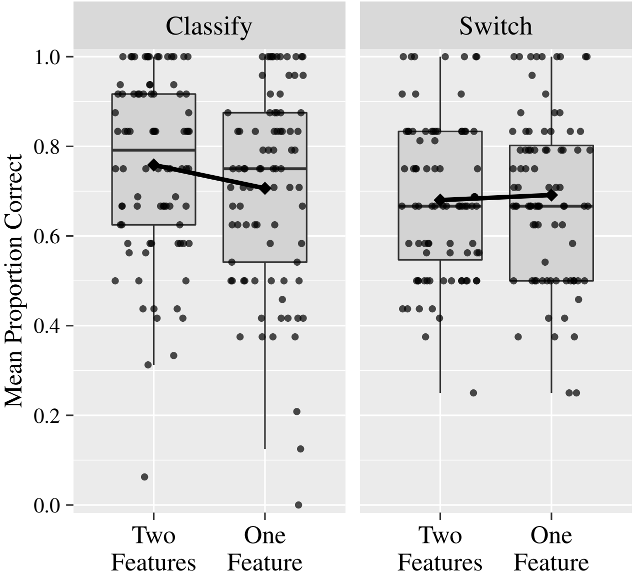


Figure 4: Mean proportion correct on the inference test phase. Tukey's boxplots present the overall accuracy pattern with black lines and diamonds indication mean differences and points representing individual means.

Training Phase Analysis. It is unclear whether the switch and classify tasks are equivalent in difficulty, prohibiting a direct comparison of training block accuracy. However, the relative difficulty of the different category structures can be examined within condition. Training accuracy was analyzed with GLMER where trial accuracy was predicted with the fixed effect of category structure and the random intercept of participant. The classify groups were not reliably different between category structures (replicating the core finding of Kurtz, Levering, Stanton, Romero, & Morris, 2013). In contrast, the switch group exhibited higher accuracy for the Type IV category structure as compared to Type II (Beta Estimate $= 0.684$, $SE = 0.24$, Wald $Z = 2.857$, $p = .004$) and Type III (Beta Estimate $= 0.687$, $SE = 0.24$, Wald $Z = 2.916$, $p = .004$) (see Figure 3).

Test Phase Analysis. To preview the test phase results, we first note that there was no reliable difference in accuracy between the learning conditions when collapsing across test phases. Breaking this down by test phase, the classify group was more accurate on inference and classification test trials and the switch group was more accurate on switch test trials. Perhaps most interestingly, learning condition and example type (partial versus full) interacted in the inference and classification tests: the switch group had higher accuracy on incomplete exemplars than on complete training examples, but the classify learners exhibited the opposite pattern. This pattern of results remains consistent when participants performing below chance are removed.

Inference Test. Inference test accuracy was analyzed as the dependent variable in a GLMER model with learning condition and example type included as interacting fixed effects and participant included as a random intercept. Overall, the switch group was less accurate than the classify group on the inference test trials (Beta Estimate $= -0.502$, $SE = 0.18$, Wald $Z = -2.736$, $p = .006$). Accuracy was worse on one-feature trials when collapsing across group (Beta Estimate $= -0.304$, $SE = 0.10$, Wald $Z = -3.08$, $p = .006$), but the interaction between learning condition and example type shows that this decrease in accuracy was not present in the switch group (Beta Estimate $= 0.38$, $SE = 0.14$, Wald $Z = 2.814$, $p = .005$) (see Figure 4).

Turning to the effect of category structure on accuracy, no reliable differences were found between category types for the classify group. The switch group, however, was more likely to be accurate if they were assigned Type IV as compared to Type II (Beta Estimate $= 0.551$, $SE = 0.24$, Wald $Z = 2.303$, $p = .02$).

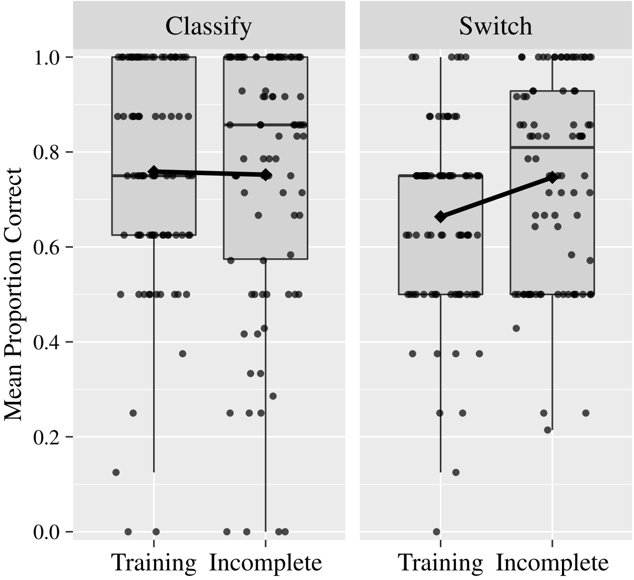


Figure 5: Mean proportion correct on the classification test phase split by condition and exemplar type.

Classification Test. The analysis approach for the classification test was similar to that of the inference test (save the dependent variable). The model uncovered reliable effects of learning condition and the interaction between learning condition and example type (See Figure 5). The switch group was less accurate overall (Beta Estimate $= -0.646$, $SE = 0.22$, Wald $Z = -2.905$, $p = .004$). However, a significant interaction shows that the switch group was more accurate on incomplete trials than on training trials (Beta Estimate $= 0.503$, $SE = 0.18$, Wald $Z = 2.722$, $p = .007$) (see Figure 5).

Category type effects were also found for classification test. The switch group was less accurate on Type II compared to Type III (Beta Estimate $= -0.725$, $SE = 0.26$, Wald $Z = 2.805$, $p = .005$) and Type IV (Beta Estimate $= -0.762$, $SE = 0.26$, Wald $Z = 2.9$, $p = .004$). No category differences were found for the classify group.

Switch Test. Accuracy on switch test phase trials was analyzed with a GLMER model with learning condition included as a fixed effect and participant included as a random intercept. The results show that the switch group was more accurate on the switch test than the classify group (Beta Estimate $= 1.683$, $SE = 0.37$, Wald $Z = 4.576$, $p < .001$). No differences between category structures were found.

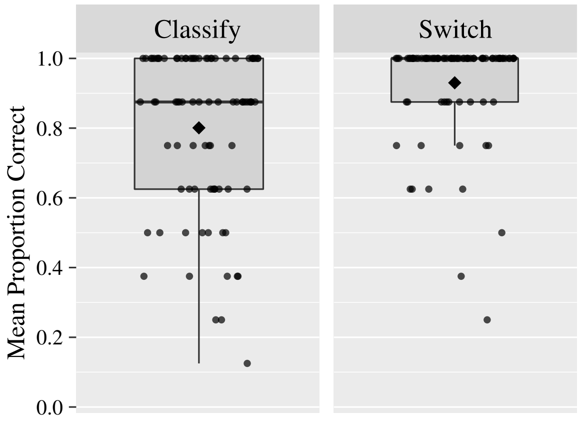
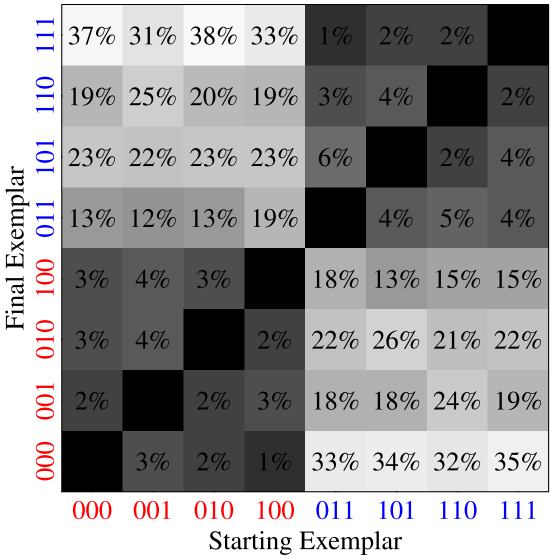


Figure 6: Mean proportion correct on the switch test phase.

Switch Test Characteristics. Given that the results have shown effects of category structure and exemplar type that are unique to the switch condition, it is of interest to characterize the switch patterns that participants used to learn the categories. These data can be compactly visualized in the style of a confusion matrix, containing the conditional probability of each possible switch, given a starting exemplar (i.e., 010 → 011). Although space does not permit a comprehensive display of this data (see the supplementary materials[1]), the switching patterns exhibited by the Type IV learners are especially illustrative (see Figure 7). These learners demonstrated a tendency to switch exemplars to the prototype of the target category. This pattern suggests that switch learners had acquired the knowledge that each class in this category structure is described by a family resemblance structure.

Discussion

We reported a novel categorization experiment exploring the differences between learning under a traditional classification learning mode and a novel switch task. Broadly speaking, the results presented in this report are consistent with known effects of category use on category learning (Markman & Ross, 2003) – learners in the switch and classify conditions not only showed advantages on their respective tasks, but also differences in their inference responses to novel items of varying completeness.



	000	001	010	100	011	101	110	111
111	37%	31%	38%	33%	1%	2%	2%	
110	19%	25%	20%	19%	3%	4%		2%
101	23%	22%	23%	23%	6%		2%	4%
011	13%	12%	13%	19%		4%	5%	4%
100	3%	4%	3%		18%	13%	15%	15%
010	3%	4%		2%	22%	26%	21%	22%
001	2%		2%	3%	18%	18%	24%	19%
000		3%	2%	1%	33%	34%	32%	35%

Figure 7: Confusion matrix depicting switch behavior during the Type IV training phase. Values represent the conditional probability of each switch, given a starting exemplar. Columns sum to 100%. Learners most commonly switched to the prototypes (000, 111).

In accounting for these effects, it is useful to consider the knowledge required to complete the switch task. Specifically, towards the goal of executing a successful class switch, learners need to understand how the present feature values relate to the target category, as well as how the current and target classes differ across the feature space. On its face, the advantage switch learners show on partial items (relative to complete items) may indicate that these individuals are better equipped to process items analytically – as a collection of parts rather than wholes. This interpretation is sensible given the nature of the switch task, where learners presumably become familiar with how elements of the exemplars relate to the class label. However, this would not explain the decline in performance on full items.

Past research on inference learning has uncovered higher (e.g., Sakamoto & Love, 2010) and lower (e.g., Sweller & Hayes, 2010) accuracy at test in relation to traditional classification. Considering the similarities between switch and inference, it is puzzling that the classification group was reliably more accurate on the inference test.

Still, there are many commonalities between inference learning outcomes and the results presented here. Switch learners had higher accuracy on family resemblance based categories as evidenced by the Type IV group's higher accuracy on the training phase and the classification and inference test phases as compared to the Type II (non-linearly separable category structure) group. Similar advantages on family resemblance categories are found for inference as well (Yamauchi, Love, & Markman, 2002).

Switch learning produced higher accuracy on exemplars with features omitted as compared to complete exemplars while the opposite was observed in the classification learning group in our study. This result is mirrored in investigations of inference learning (Anderson, Ross, & Chin-Parker, 2002).

Another result of switch learning is that the family resemblance category structure was learned quite quickly. By evaluating the switch behavior during training (Figure 7) it can be seen that the prototypes are favored during this process – another commonality that can be tied back to a distinction made between inference and classification (Johansen & Kruschke, 2005; Yamauchi & Markman, 1998). Furthermore, the relative ease of learning across the category structures under the switch learning mode provides new evidence on the variable nature of the ordering of acquisition of the SHJ categories (Kurtz et al., 2013).

Given that real-world categories often conform to a family resemblance structure (Rosch & Mervis, 1975), future work will explore the sensitivity to within-category regularities and rapid learning of Type IV seen with switch-based learning.

Acknowledgments

We thank Kimery Levering for developing the stimuli used in this study, Sarah Laszlo and the members of the Binghamton University Modeling Meeting, and the members of the LaRC (Learning and Representation in Cognition) lab.

References

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

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

Johansen, M. K., & Kruschke, J. K. (2005). Category representation for classification and feature inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1433–1458.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*(4), 560–576.

Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. *Psychology of Learning and Motivation*, *63*, 77–114.

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), 552–572.

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & cognition*, *43*(2), 266–282.

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

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

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge: MIT press.

Murphy, G. L. (2005). The study of concepts inside and outside the laboratory: Medin versus medin. In W. Ahn, R. Goldstone, B. Love, A. Markman, & P. Wolff (Eds.), *Categorization inside and outside the lab: Essays in honor of Douglas L. Medin* (pp. 179–195). Washington, DC: American Psychological Association.

Peirce, J. W. (2007). Psychopy – psychophysics software in python. *Journal of Neuroscience Methods*, *162*(1), 8–13.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

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

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Sakamoto, Y., & Love, B. C. (2010). Learning and retention through predictive inference and classification. *Journal of Experimental Psychology: Applied*, *16*, 361–377.

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

Solomon, K. O., Medin, D. L., & Lynch, E. (1999). Concepts do more than categorize. *Trends in cognitive sciences*, *3*(3), 99–105.

Sweller, N., & Hayes, B. K. (2010). More than one kind of inference: Re-examining what's learned in feature inference and classification. *The Quarterly Journal of Experimental Psychology*, *63*(8), 1568–1589.

Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 585.

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