



# Revisiting the linear separability constraint: New implications for theories of human category learning

Kimery R. Levering<sup>1</sup> · Nolan Conaway<sup>2</sup> · Kenneth J. Kurtz<sup>2</sup>

© The Psychonomic Society, Inc. 2019

## Abstract

While the ability to acquire non-linearly separable (NLS) classifications is well documented in the study of human category learning, the relative ease of learning compared to a linear separable structure is difficult to evaluate without potential confounds. Medin and Schwanenflugel (*Journal of Experimental Psychology: Human Learning and Memory*, 7, 355-368, 1981) were the first to demonstrate that NLS classifications are not more difficult to acquire than linearly separable ones when structures are equated in terms of within- and between-category similarities. However, their evidence is less sturdy than might be expected due to non-standard methodology and low sample size. We conducted a conceptual replication to clarify the behavioral picture and perform qualitative testing of formal models. The behavioral results not only showed a lack of advantage for the linearly separable (LS) structure, but revealed a stronger finding: the NLS structure was reliably easier to acquire. Differences in the relative ease of NLS learners to master certain items yielded evidence for the existence of distinct learner subgroups, one marked by significantly easier (not harder) learning of exception items. Comparing the qualitative fits of leading computational models to the human learning performance confirmed that a pure prototype account, even with contemporary updates, remains incompatible with the data. However, exemplar models and similarity-based models grounded in sophisticated forms of abstraction-based learning successfully account for the NLS advantage. In sum, evidence against a linear separability constraint is redoubled, and the observed NLS advantage along with behavioral patterns seen at the subgroup and item level provide a valuable basis for comprehensive evaluation of competing theoretical accounts and models.

**Keywords** Concept learning · Categories · Linearly separable · Connectionist modeling · Representation

## Introduction

A longstanding tradition in the category-learning literature is to compare human learning of different types of category structures in order to understand how concepts are acquired and organized. Behavioral studies of what types of categories are easier or harder to learn serve to support or falsify theories and models of the processes and representations underlying the psychology of categorization. A commonly revisited question is the extent to which linear separability is a constraint on concept learning and organization. Though data bearing on this question were already in existence (e.g., Shepard,

Hovland, & Jenkins, 1961), Medin and Schwanenflugel's study (1981) was the first investigation to find evidence against a linear constraint after including appropriate controls for item similarity. Despite being often cited and used as a benchmark data set for testing computational models, there have been concerns about methodological factors in this research report including their use of non-standard materials, low power, and sub-mastery levels of learning. In this paper, we revisit an experiment from Medin and Schwanenflugel (1981) in an attempt to collect a more conclusive data set that clarifies the core phenomenon and provides a solid basis from which to evaluate the explanatory success of leading similarity-based models.

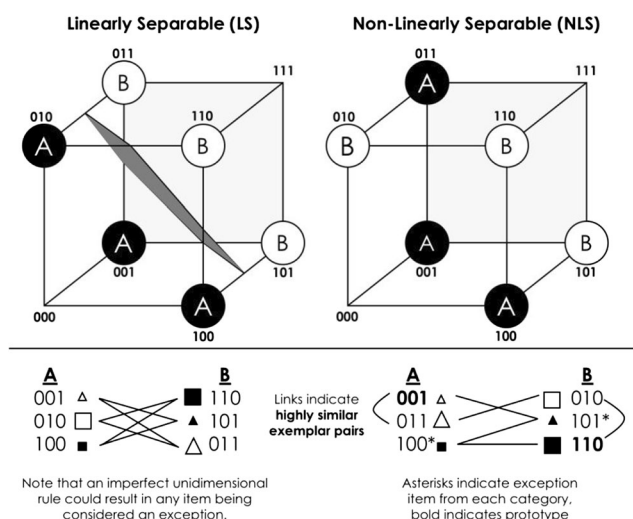
✉ Kimery R. Levering  
kimery.levering@marist.edu

<sup>1</sup> Department of Psychology, Marist College,  
Poughkeepsie, NY 12601, USA

<sup>2</sup> Department of Psychology, Binghamton University,  
Binghamton, NY, USA

## Linear separability

Linearly separable categories (see Fig. 1) occupy distinct regions of multidimensional stimulus space such that they can be fully partitioned with a single linear decision bound (hypersurface). Accordingly, class decisions can be computed as



**Fig. 1** Linearly (LS) and non-linearly separable (NLS) category structures used in Medin and Schwanenflugel (1981, Exp. 4) and the present study. The three members of each category varied in shape (square or triangle), size (1.5 in. or .75 in.), and shading (black or white). Members are represented here as vertices of a three-dimensional cube, as rows of binary digits, and as one possible instantiation of assignment to geometric shapes used in the current study. In terms of additive similarity, LS categories contained no high similarity within-category exemplar pairs and six high similarity between-category pairs, while NLS included two high similarity within-category exemplar pairs and only four high similarity between-category pairs. Both category problems had equivalent within- and between-category similarity and structure ratios of 1.25

a function of a weighted, linear combination of feature information. Non-linearly separable categories cannot be partitioned using a single linear decision surface without at least one category example being on the wrong side. If humans rely on linear separability as a primary means of differentiating between natural categories, then non-linearly separable categories should be more difficult or impossible to learn. This does not seem to be the case. For example, a well-known case of a non-linearly separable structure is the exclusive-OR (XOR) problem, which is highly amenable to human learning and sometimes quite rapidly acquired (Kurtz, Levering, Stanton, Romero, & Morris, 2013; Nosofsky, Palmeri, & McKinley, 1994; Shepard et al., 1961; see also Medin, Altom, Edelson, & Freko, 1982).

Theories and models of category learning differ in their reliance on linear separability as a primary constraint on category learning. On one end of the spectrum are independent-cue models, which assume category judgments to be made by summing information (e.g., similarity, distance, cue validity, associative frequency) from component cue dimensions. The most well-established class of such models is based on prototype theory (e.g., Posner & Keele, 1968; Reed, 1972; Rosch & Mervis, 1975) and formalized by considering the overall distance from a computed average (mean or mode) based on prior exposure to category examples. Because independent-cue models assume that additive rules guide human categorization, they predict category structures organized accordingly

(e.g., linearly separable, LS, category structures) will be learned more quickly than non-linearly separable (NLS), which may not be learned at all.

Alternatively, relational coding (or interactive cue) models assume learning to be unconstrained by linear separability. Most notably, exemplar models (e.g., Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986) consider classification decisions to be based *not* on summed evidence from independent dimensions but rather on summed similarity to entire examples (or a subset of attended features) – computed using a multiplicative rule that combines evidence along each dimension. Unlike prototype models, these models need not rely on any category-level knowledge about cue regularity. Despite this fundamental difference in model design, exemplar models can successfully account for learning phenomena thought to be strong evidence for prototype theory – notably the finding that examples representing central values along relevant dimensions are often acquired more quickly, classified more accurately, and forgotten more slowly. Exemplar models achieve this result because the multiplicative rule results in the elevated weight of high exemplar similarity (between or within categories) on classification decisions and exemplars with central values on relevant dimensions often bear great resemblance to other trained examples in their category. However, when high exemplar-exemplar similarity within categories (and not between categories) is pitted against linear separability, exemplar models predict that linearly separable categories would be learned more slowly (see Medin & Schaffer, 1978, for more detail).

### A classic study of linear separability

Medin and Schwanenflugel (1981) performed a critical test of this difference between the leading theoretical accounts of the time. In a series of four experiments, they compared learning of linearly separable and non-linearly separable categories across a number of combinations of within- and between-category similarity, and a number of highly similar items. The details of these combinations can be seen in Fig. 2. In categories with binary feature values, within-category additive similarity is typically measured as the average number of features shared between members of the same category and indicates how clustered together examples in each category are. Between-category additive similarity is measured as the average number of features shared between members of different categories and indicates how distinct categories are from each other. Structure ratios are the proportion when within-category similarity is divided by between-category similarity – this provides a standard way of measuring the overall coherence of a category structure (Homa, Rhoads, & Chambliss, 1979; Smith, Murray, & Minda, 1997). A structure ratio of 1 indicates that examples within each category are equally similar to each other as they are to members of a different category (no

category coherence) while a structure ratio of 2 indicates strong family resemblance within categories. Matching category structures with regard to these measures of additive similarity is critical because, without this, any claim about differences in learning between an LS and an NLS structure could be attributable to qualities of well formedness of the structures rather than their linear-separability status. In Experiments 3 and 4 of Medin and Schwanenflugel's (1981) study, LS and NLS structures were perfectly matched on average additive exemplar similarity within and between categories and, therefore, the resulting structure ratios. Because of this careful attention to controlling these similarity factors, Experiments 3 and 4 have been cited and formally modeled frequently in the literature (e.g., Kruschke, 1992; Kruschke, 1993; Nosofsky et al., 1994; Smith et al., 1997). In the current study, we use the structures from Experiment 4 because, as noted in the original paper, the LS category structure represents the simplest case of summation (two-out-of-three rule).

Based on structural differences, prototype and exemplar models make opposite predictions for the relative ease of learning the categories shown in Fig. 1. Prototype models predict the LS category structure to be more easily learned because summing evidence along each dimension would be a successful strategy for classification in this case, but not for the NLS category. Exemplar models predict the NLS category structure to be more easily learned on the basis of having a greater number of highly similar within-category pairs and a smaller number of highly similar between-category pairs (as seen in Fig. 1). Because the category structures are matched in terms of within- and between-category additive similarity and their structure ratios, overall coherence or distinctiveness of the categories would not lead to an advantage for learners of either structure.

In Experiments 3 and 4, Medin and Schwanenflugel (1981) found that the mean number of errors for LS and NLS were not significantly different from each other.<sup>1</sup> Despite a lack of a statistically reliable difference, the result that LS learners did not perform better than NLS learners directly contradicted the prototype view and provided strong support for Medin and Schaffer's (1978) context model. More recently, this non-difference has been used to criticize models like standard back-propagation as being overly sensitive to linear boundaries (Kruschke, 1992, 1993) and has helped to bolster the case for formal models like the rational model (Anderson, 1991), RULEX (Nosofsky et al., 1994), the DIVERgent Autoencoder (DIVA: Kurtz, 2007), the generalized version of the context

model (GCM: Nosofsky, 1986), and its connectionist implementation (ALCOVE: Kruschke, 1992).

While the findings were influential, the conclusions from the experiment are subject to a number of criticisms. First, the number of subjects was quite small (16 per group in Experiment 4) and therefore the null result found could have been due to lack of power. Second, there was a low level of mastery demonstrated by the participants as seen by the fact that only 11/32 reached a learning criterion of two errorless blocks across 18 blocks of learning in Experiment 4. As noted by critics (Blair & Homa, 2001; Smith et al., 1997), sub-mastery levels of aggregate performance in this case allow for a variety of possible explanations about representation and/or strategy. For example, learners could have been applying an imperfect linear decision rule – a simple unidimensional rule on any dimension in the NLS structure would correctly classify 66% of the examples. To that point: the overall proportion correct of NLS learners was only .65. In sum, a failure to find a statistically significant difference using these structures does not rule out the possibility of a linearly separable constraint – despite the fact that it has been regularly taken to mean this. In addition, due to the low levels of learning and lack of more sensitive measures, the original study does not offer much opportunity to evaluate learner strategy and representation. We pursue these questions in the current study by increasing the power, providing more training, and adding a test phase that includes typicality ratings.

A further concern with the original study is that the materials were idiosyncratic in ways that limit the suitability for model fitting. Specifically, in Experiments 3 and 4, stimuli were photographs of faces differing in hair color (dark, light), hair length (long, short), and smile type (closed, open), and different instantiations of the logical dimensions were presented for each of the 18 blocks (i.e., the same photo was never shown twice). While this type of stimuli may be more ecologically valid, it is difficult to ensure that it is consistent with the separable and invariant encoding of feature dimensions assumed by models. For example, two of the dimensions used (hair color and hair length) are elements of the same physical feature, and while the smile type is physically separable, there is robust evidence that elements of faces such as mouths are processed holistically rather than in terms of separate dimensions (see Maurer, Le Grand, & Mondloch, 2002). While geometric shapes traditional to artificial category learning experiments were used in the first two experiments, these experiments did not provide the critical controls on item similarity.

## The current study

Because the results of Medin and Schwanenflugel (1981) have been used so frequently to justify constraints on models and have so substantially impacted theoretical development in the category learning literature (e.g., Anderson, 1991; Kruschke,

<sup>1</sup> In Experiment 2, NLS learners committed significantly less errors than LS learners. However, as can be seen in Fig. 2, these participants were learning a category distinction that had a higher structure ratio than the LS learners and it is possible performance was improved because of this.

Category	EXPERIMENT 1				EXPERIMENT 2				EXPERIMENT 3				EXPERIMENT 4			
	LS		NLS		LS		NLS		LS		NLS		LS		NLS	
	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
Membership	1011	1001	1000	0110	1110	1010	1000	0001	0111	1000	1100	0000	011	100	100	000
	1010	0010	0111	1001	1011	0110	1010	0100	1110	0001	0011	0101	110	001	011	001
	1101	0100	1110	0000	1101	0001	1111	1011	1001	0110	1111	1010	101	010	111	110
	0110	0001	1011	0001	0111	1100	0111	0000								
Average Similarity	Within: 2.31 Between: 1.75		Within: 2.25 Between: 1.75		Within: 2.31 Between: 1.88		Within: 2.31 Between: 1.69		Within: 2.22 Between: 1.78		Within: 2.22 Between: 1.78		Within: 1.67 Between: 1.33		Within: 1.67 Between: 1.33	
Num. of High Similarity Items	Within: 2 Between: 5		Within: 2 Between: 5		Within: 0 Between: 6		Within: 4 Between: 3		Within: 0 Between: 4		Within: 0 Between: 0		Within: 0 Between: 6		Within: 2 Between: 4	
Structure Ratio	1.32		1.29		1.23		1.37		1.25		1.25		1.25		1.25	
Stimuli	<b>Geometric figures:</b> Form (triangle, circle) Size (small, large) Color (red, blue) Position (one or two)				<b>Geometric figures:</b> Form (triangle, circle) Size (small, large) Color (red, blue) Position (one or two)				<b>Yearbook photos of women*:</b> Hair color (light, dark) Hair length (long, short) Shirt color (light, dark) Smile (open, closed)				<b>Yearbook photos of women*:</b> Hair color (light, dark) Hair length (long, short) Smile (open, closed)			
Sample Size	N = 64				N = 64				N = 64				N = 32			
Learning	<b>Total Blocks:</b> 16 <b>Criterion:</b> 1 perfect block				<b>Total Blocks:</b> 20 <b>Criterion:</b> 1 perfect block				<b>Total Blocks:</b> 18 <b>Criterion:</b> 1 perfect block				<b>Total Blocks:</b> 18 <b>Criterion:</b> 2 perfect blocks			
Results	Errors: NLS = LS				Errors: NLS < LS				Errors: NLS = LS				Errors: NLS = LS			

\* Each example had a unique instantiation of feature values

Fig. 2 Overview of experiments from Medin and Schwanenflugel (1981)

1992; Kruschke, 1993; Kurtz, 2007; Nosofsky, 1986), it is important to re-examine the findings to address the aforementioned concerns. Our study is a conceptual replication using the category structures of Medin and Schwanenflugel (1981, Exp. 4) with more standard materials (geometric figures), longer training duration (25 blocks instead of 18), and much higher power. This experiment is expected to yield more definitive evidence as to whether the classic demonstration of LS/NLS equivalence accurately captures human performance on these tasks. The addition of a test phase (classification with no corrective feedback followed by a typicality rating) allows for a more nuanced investigation of learner strategies and representations.

Above and beyond working to better specify the behavioral benchmark, the world of formal models of category learning has grown richer in the intervening years, so it is critical to update the theoretical implications of the observed learning performance. In the modeling component of the present research, we simulate the time course of learning the LS and NLS structures using four models: a canonical exemplar model (ALCOVE), a comparably-outfitted model using prototypes instead of exemplar reference points, a more sophisticated reference point model based on clusters that adaptively take the form of either prototypes, sub-prototypes or exemplars (SUSTAIN; Love, Medin, & Gureckis, 2004), and a connectionist model (DIVERgent Autoencoder, DIVA; Kurtz, 2007, 2015) that uses autoassociative, error-driven learning to estimate a generative model of the regularities of each category. See Fig. 3 for key differences between models.

The latter account (DIVA) is of unique interest as it explains category learning from outside of the reference

point framework. DIVA uses the match between the input features and the expected features (construal) with respect to each category at the output layer to predict the likelihood of category membership. The reconstructive success (i.e., the ability to recover the original features after projection into a learned recoding space at the shared hidden layer) is measured using a sum-squared error metric for each category, which is submitted to a Luce choice rule to produce probabilistic responding. Unlike models that predict slower learning depending on how difficult it is to position a successful classification boundary in either input space or recoded space (i.e., projections of the input into a multidimensional space at an intermediate layer) – DIVA learns to classify more slowly insofar as it is difficult to discover recodings of the category members that allow successful prediction (reconstruction) of the input features. This can be thought of as follows: categorization tasks are harder to learn to the extent that dissimilar category members must be represented similarly in recoding space.

## Method

### Participants

Two hundred and seventy Marist College students participated in exchange for partial fulfillment of course credit. 144 participants were randomly assigned to the linearly separable category structure and 126 participants were assigned to the non-linearly separable category structure.



Model	Recoding	Prediction
ALCOVE	Shepard similarity to exemplars	Association to classes
SUSTAIN	Shepard similarity to clusters	Association to classes
PROTO-ALCOVE	Shepard similarity to each prototype	Association to classes
DIVA	Error-driven recoding space	Feature values with respect to each class

Fig. 3 Key design principles of the four models used in the present study

## Stimuli and category structures

Stimuli were geometric shapes varying along the following three binary dimensions: shape (square or triangle), size (1.5 in. or .75 in.), and shading (black or white). While all combinations of these features create a total of eight examples, categories were made up of only six of these (three examples per category, see Fig. 1). Logical structure of categories was consistent with the linearly and nonlinearly separable categories of Medin and Schwanenflugel (1981, Exp. 4), and assignment of physical features to logical structure was completely counterbalanced. Both category problems had equivalent within- and between-category additive similarity and structure ratios of 1.25. However, the linearly separable category contained no high similarity within-category exemplar pairs (and six high similarity between-category pairs), while the nonlinearly separable category included two high similarity within-category exemplar pairs (and only four high similarity between-category pairs).

## Procedure

**Training phase** On each self-paced trial, a randomly selected example was presented and participants were instructed to decide which of two categories (Alpha or Beta) the example belonged to. After their response, feedback about correct category membership was given. Participants were trained on 25 blocks of classification learning, each block consisting of the classification of all six examples. Unlike Medin and Schwanenflugel (1981), no learning criterion was set to indicate mastery and trigger the assumption of future correct trials. The more conservative decision to not use a learning criterion reduces the risk of treating a subject as having reached mastery when they had by chance been performing above their level of knowledge.

**Test phase** After completing the learning phase, a test phase consisted of classification in randomized order with no corrective feedback of the six training examples plus the two untrained examples (000, 111) that complete the set of possible items for three binary-valued dimensions. On each test trial, after a classification response was given, participants indicated how typical each example was of the category on a scale of 1 (not typical) to 9 (highly typical).

## Results and discussion

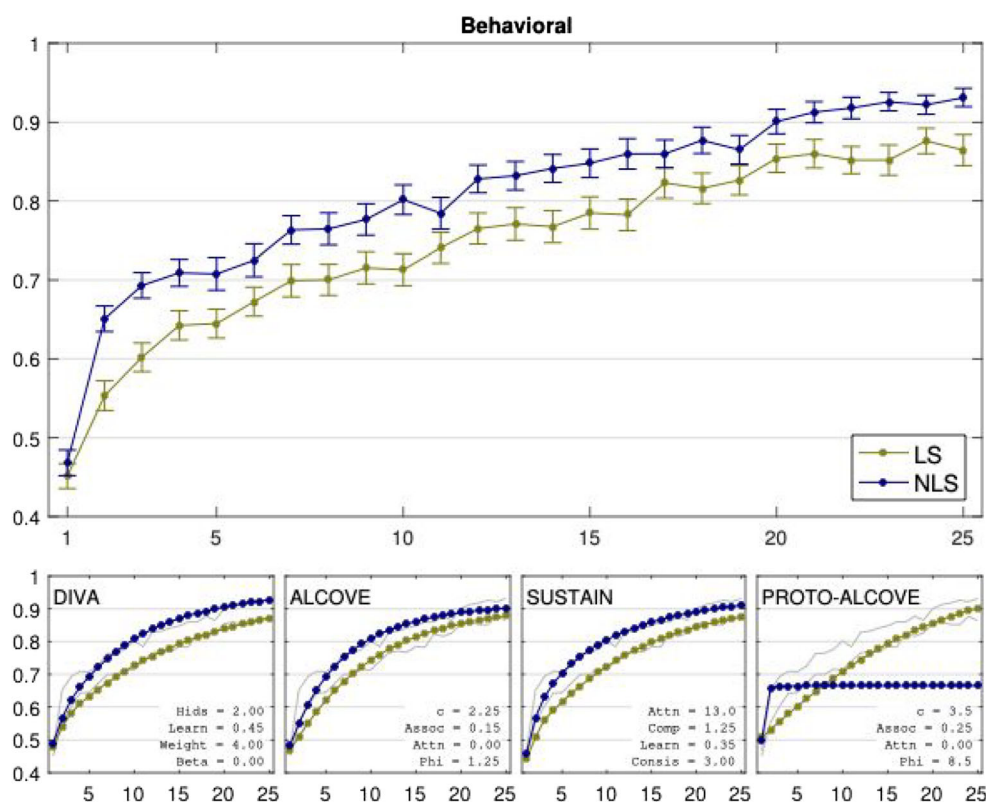
### Behavioral data<sup>2</sup>

#### Aggregate learning and test performance for trained items

Performance across learning for the two classification problems can be seen in Fig. 4. A 2 (LS, NLS) by 25 (learning blocks) mixed ANOVA was conducted to explore differences in learning performance between the classification conditions. Unsurprisingly, there was a significant main effect of learning trial,  $F(24, 6432) = 119.768, p < .001, \eta^2 = .309$ , indicating that accuracy increased over learning trials. However, in contrast to the original findings, there was also a significant main effect of classification problem,  $F(1, 268) = 15.054, p < .001, \eta^2 = .053$ . Specifically, proportion correct for the NLS problem ( $M = .806, SD = .125$ ) was significantly greater than for the LS problem ( $M = .745, SD = .134$ ). A lack of interaction between the two variables,  $p = .698, \eta^2 = .003$ , indicated that this difference in performance was consistent across learning. This difference was also observed at test, where performance of the NLS group ( $M = .923, SD = .149$ ) was significantly higher than that of the LS group ( $M = .838, SD = .201$ ),  $t(264) = 3.902, p < .001, d = .453$ . With regard to our aim of increasing power, we note that with an effect of this size, the original experiment would have had only 26.5% power, as opposed to the 97.2% observed in the current study.

**Representation of NLS categories** Patterns of classification performance and typicality ratings across items may help to clarify underlying category representations or reveal distinct types of learners. Of particular interest in the existing literature is performance on NLS exception items. These items (A100 and B101, shown with asterisks in Fig. 1) are on opposite sides of a prototype-based linear boundary for the NLS structure and therefore would be incorrectly classified according to this strategy. These items are also least similar to members of their own category and most similar to members of the other category, so the exemplar view also predicts more difficult learning of these items. Consistent with previous research and these predictions, NLS exception items showed significantly worse performance ( $M = .755, SD = .159$ ) than the other items ( $M = .832, SD = .128$ ) during training,  $t(125) = 6.71, p < .001$ . The fact that the NLS exception items can be learned at all has been taken as evidence against a linear constraint and, more specifically, against prototype models that classify these items into the wrong category. Results from the current experiment serve to further highlight the learnability of these exception items. At test, performance on the NLS exception items was not reliably worse than the other items,  $t(125) = 1.62, p = .107$ . In fact, only 18/126 participants made even one mistake on an exception item during the test phase

<sup>2</sup> Raw data files can be found at <https://osf.io/x3gqe>



**Fig. 4** Proportion correct across blocks of learning based on human behavioral data and model fits for DIVA, ALCOVE, SUSTAIN, and ALCOVE outfitted for prototype representation. Parameters associated with best fits are listed for each model

and only three participants put both exception items into the incorrect category (as predicted based on a linear boundary). In sum, these data provide powerful evidence against a linearly separable constraint.

Another characteristic of the NLS structure is that one item in each category (bolded items A001 and B110 in Fig. 1) possesses the most common features for the category on each dimension and has the highest overall similarity to members of its own group and lowest similarity to members of the opposite category. Based on its proximity to the central tendency of its category (and not any assumptions of psychological representation), we follow tradition by referring to this item as the prototype. These items were classified more accurately than intermediate items during training,  $t(125) = 3.696$ ,  $p < .001$ , and at test,  $t(125) = 1.999$ ,  $p = .048$ . They were not, however, rated more typical at test,  $p = .145$ . Generalization performance on the untrained items at test (000 and 111) varied widely. If classification based on similarity to exemplars or prototypes is treated as the correct response (i.e., A000 and B111), then performance was quite low ( $M = .56$ ,  $SD = .38$ ) but reliably better than chance,  $t(125) = 16.383$ ,  $p < .001$ .

As noted by other researchers (e.g., Blair & Homa, 2001; Smith et al., 1997), it is possible for aggregated data to hide the presence of multiple learner profiles with distinct patterns of performance. In the past, “prototype” learners have been characterized as those who persist in putting exception items

into the wrong category despite feedback indicating otherwise. Because our participants were making so few errors by the end of training, very few, if any, participants could be characterized as purely “prototype” learners. However, there are a number of possible strategies/representations that could explain significantly poorer performance on exception items. We also note that 28 participants actually performed *better* on exception items during training compared to the average of the other items. This is further evidence against a linear constraint and also suggests that these learners may have acquired a different representation of the category – one that is not well captured by standard explanations.

In an effort to understand possible learner subgroups, we divided participants into three types based on the clearest differentiating signature in the data: how well they learned NLS exception items during training. To minimize noise based on overall performance, we created a difference score by subtracting each participant’s average performance on non-exception items from their performance on exception items. A difference score of zero indicates that the participant’s performance on exception items was no different from their performance on the other four items in the category. To the extent that the difference score is negative, participants had trouble learning the exception item relative to other items in the category. A positive difference score indicates that participants demonstrated higher accuracy on exception items across learning. This

variable was normally distributed (see Fig. 5) and shifted toward negative difference scores, reflecting the prevalence of learners who learned exception items more slowly than other items. Despite a normal distribution which suggests no clear qualitative differences, it is possible that the difference score distribution reflects meaningful subgroups. In particular, we were curious about the subset of learners who performed *better* on exception items during training. Accordingly, we classified learners as members of the High Exception (HighX) subgroup when their difference score was more than half of a standard deviation *above* the mean difference score ( $M = .056$ ,  $SD = .079$ ) or members of the Low Exception-Strong (LowX-Strong) subgroup when their difference scores were more than half a standard deviation *below* the mean difference score ( $M = -.243$ ,  $SD = .089$ ). Those in the middle were considered Low Exception-weak (LowX-Weak) learners ( $M = -.074$ ,  $SD = .035$ ) because all of these participants performed worse on exception items during training, but to a lesser degree than the LowX-Strong learners. The patterns of performance reported below are consistent across a variety of cutoff points (from .25 to 1 standard deviations).

We tested whether the identified subgroups differed in their learning of the classification problem (see Fig. 6). A 3 (subgroup)  $\times$  5 (learning quintile) repeated-measures ANOVA revealed a significant main effect of subgroup,  $F(2,125) = 7.436$ ,  $p = .001$ , and no interaction,  $p = .164$ . *Post hoc* tests showed that the LowX-Weak group had higher proportion correct than both the LowX-Strong,  $p = .001$ , and HighX,  $p = .026$ , groups. Performance differences between the two LowX groups can be explained simply by the fact that LowX-strong learners by definition had a particularly hard time with exception items. It makes sense that this increased difficulty is reflected in overall learning accuracy because performance on the other items did not differ significantly from the LowX-weak group. Those who did not particularly struggle to learn the exceptions (HighX group) had significantly lower performance on other items, bringing their overall learning accuracy down. However, performance differences between groups were not seen in the test data,  $p = .128$ . Similarly, the groups did not differ significantly in their performance on the new items at test,  $p = .279$ .

For insight into representation, we also compared the pattern of learning, test, and typicality ratings (of correctly classified examples) across item types. See Fig. 5 for the details of these analyses. While the LowX-Strong and LowX-Weak groups showed patterns of performance across items that closely matched the aggregate data, the HighX group not only learned the exception items most easily but also did not consistently rate any type of item more typical at test. In the following section, we address possible accounts of the majority behavior (and then possible explanations of the distinctive HighX subgroup). For the majority of learners (the LowX-Strong and LowX-Weak groups), exception items were

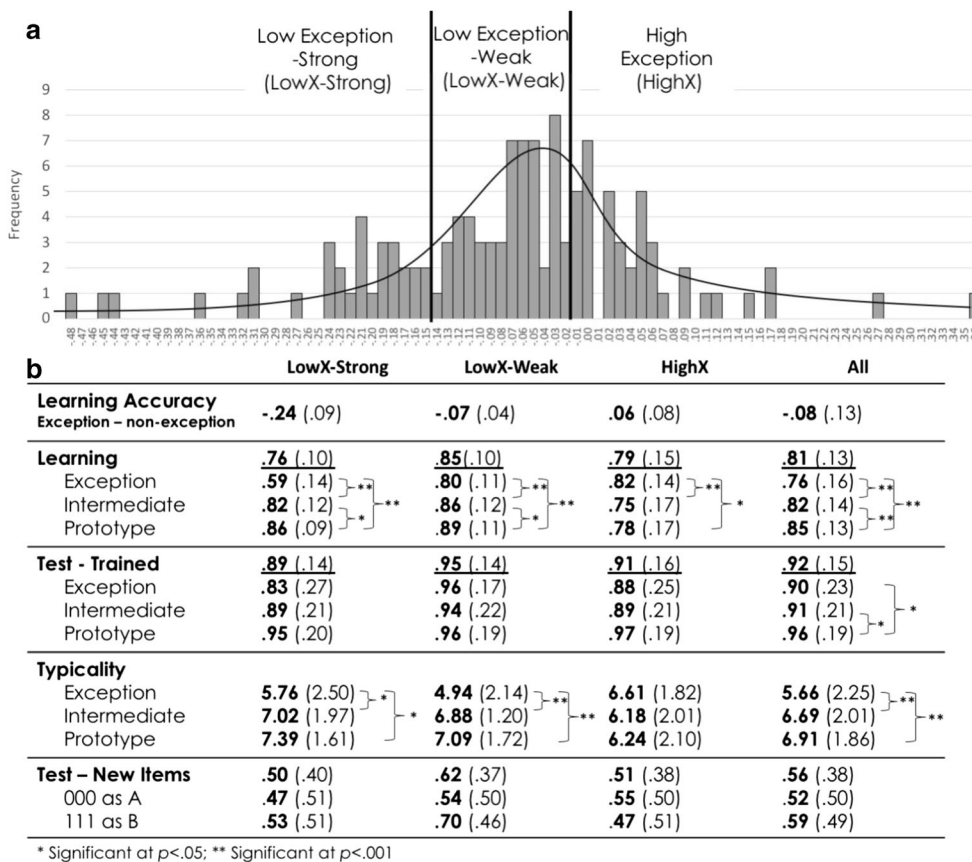
hardest to acquire and judged as less typical after mastery. The most prototypical item was learned fastest for these groups, although typicality ratings were not consistently higher.

A majority of learners in the LowX groups (21/32 participants in the LowX-Strong and 31/56 participants in the LowX-Weak group) made fewer errors on prototype items than intermediate items during learning. This pattern of performance has several possible interpretations. First, participants could be adopting an exemplar representation and gradually learning the collective associations (via stimulus generalization) between examples and their labels. On this view, the speed of learning for individual items is predictable based on the degree of similarity to other items in the category and the degree of difference from items in the other category (e.g., the exception items would be learned less well due to being more like members of the other category than members of its own).

It is also possible that participants adopt a unidimensional rule-plus-exception (Nosofsky et al., 1994) representation. For example, using the feature assignment in Fig. 1, they could learn that Category A members are triangles and Category B members are squares and then memorize the two exceptions to this rule. Of course, learners would be equally likely to adopt a unidimensional rule on any of the three dimensions. Accordingly, when averaged across learners adopting this approach for each of the three dimensions, the expected pattern is consistent with the current findings because the harder-to-learn items inconsistent with the unidimensional rule align two-thirds of the time with the NLS exception and never align with the NLS prototype.

To evaluate the possibility that LowX learners are adopting a rule-plus-exception strategy, we conducted chi-squares for each participant based on the pattern of training errors for rule-consistent and rule-inconsistent items according to the three possible unidimensional rules. Using this metric, 56/88 participants performed in a manner that was consistent with a rule-plus-exception strategy, 33 of them showing a pattern consistent with a rule on the third dimension and 11 and 12 showing a pattern that was uniquely consistent with a rule on either the second or third dimensions, respectively,  $ps < .05$ . We take this as evidence that a substantial proportion of LowX learners are not adopting this strategy. Further, because equal numbers of participants would be expected to adopt rules along the three dimensions, these data suggest that a portion of those with performance consistent with a rule on the third dimension may be operating according to a different strategy. It is worth nothing that in comparison, only 8/38 participants in the HighX subgroup showed patterns of performance consistent with a rule-plus-exception strategy on any of the three dimensions.

From the perspective of similarity-based approaches that learn through abstraction, there are further explanations to consider. While a view reliant exclusively on a linear

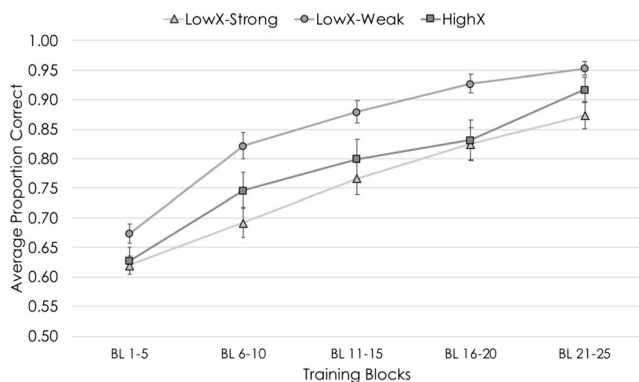


**Fig. 5** (A) Frequency distribution for learning accuracy difference score. Difference score is proportion correct for non-exception items subtracted from proportion correct for exception items. (B) Means and SDs for

performance on learning, test, and typicality, presented by subgroup and overall average for the non-linearly separable (NLS) category

boundary could not explain high performance on the exception items, participants could be abstracting a prototype based on independent additive rules along each dimension (linear boundary) and then learning the exception items through an additional mechanism such as all-or-none memorization. This view would correctly predict worst performance on the exception items because they are more similar to the prototype of the opposite category. This view would also correctly predict higher accuracy for the prototype items during learning and at

test – a pattern seen in a majority of LowX participants. More sophisticated similarity-based models (described below) that employ dynamic representation learning rather than taking the category representation from the training set suggest different accounts in terms of their design principles (i.e., SUSTAIN forming optimal clusters for class prediction; DIVA forming optimal item recodings for within-category feature prediction). In sum, the majority profile for the NLS exception items is at least partially consistent with a variety of similarity-based explanatory approaches – ongoing work based on quantitative model fitting should reveal which underlying mechanisms provide the best account.



**Fig. 6** Training performance for non-linearly separable (NLS) learner subgroups

Participants in the HighX group either find the exception items easier to learn ( $n=27$ ) or their performance on exception items was indistinguishable from other items ( $n=11$ ). At test, unlike other participants, they do not find any of the item types more typical than any others. To our knowledge, this pattern lies outside of the explanatory scope of exemplar, prototype, and rule-based approaches. One possibility is that this comes as a result of all-or-none memorization of item-label pairs (without stimulus generalization) under the particular circumstance that the exception items happen to be the first items learned. As a result of being learned first, overall learning



accuracy would be best. According to this view, exception items – or any other type of item – were not considered more typical than any other because they were just learning simple associations to a label (this strategy would afford no basis to judge typicality because there is no category representation).<sup>3</sup>

This interpretation of the performance of a distinct minority of participants would be consistent with concerns raised about small exemplar sets and impoverished category structures (see Blair & Homa, 2001, 2003; Smith & Minda, 1998; Smith et al., 1997). The root of this criticism is that participants may achieve classification success through more of an individual identification approach, i.e., examples memorized in an all-or-none fashion with no coherent category knowledge obtained. Despite the evidence of a non-trivial number of such learners, we see robust evidence of a more coherent representation from the vast majority of participants. In both the LowX-Strong and LowX-Weak groups of the NLS structure, participants are consistently learning certain items (those predicted by category learning models) better than others, and ratings of typicality reflect those differences in learning. The fact that we see differences between LS and NLS alone is enough to make clear that all-or-none memorization cannot be the main principle underlying acquisition of 3-3 categories. While all-or-none memorization of items with no stimulus generalization may have been a valid account of the non-difference between the category structures reported in the original study, it cannot explain the robust NLS advantage and the systematically differentiated patterns across items observed in the present data.

**Representation of LS categories** No significant differences were observed during learning or test between trained LS items. This is not surprising given the following: each trained example is equal in terms of between- and within-item additive similarity to examples and prototypes, an imperfect unidimensional rule is equally likely across dimensions, and usefulness of correlations between features does not differ. For the LS category structure, the untrained examples presented at test are the prototypes of each category (they contain the most common values along each feature). Despite this, learners were no more likely to classify the prototypes into their own category than the opposite category. Specifically, the overall likelihood of classifying untrained prototypes to the category most similar to trained exemplars ( $M = .537$ ,  $SD = .42$ ) was not significantly different from chance,  $t(143) = .93$ ,  $p = .30$ . The only consistent pattern was that participants were more likely to assign the new items to different categories (55/144 assigned 000 to category A and 111 to category B, 45/144

assigned 000 to category B and 111 to category A) than they were to assign them both to the same category (20/144 classified them both as As, 24/144 classified them both as Bs). This pattern (and chance performance) held when the analysis was limited to only participants who had made no errors classifying the trained examples at test ( $n = 67$ , 47%),  $t(66) = 1.40$ ,  $p = .16$ . As such, prototype-, exemplar-, and rule-based accounts make a prediction that is not borne out in the present data.

**Summary of behavioral results** Results show a clear NLS advantage in the ease of learning. For NLS learners, a majority of participants found the exception item harder to learn and less typical than other items. Items possessing the most common features were easier to learn and were mastered most successfully. A subset of learners found the exception items easier to learn and showed no clear typicality pattern. For LS learners, there were no discernable differences in learning strategy or representation based on item-level performance, but classification of new items indicated a lack of generalization to untrained prototypes.

These results provide clear evidence against a linear separability constraint given the materials and category structures used in this study. We note that this result may not generalize to variation in aspects of category structure, dimensions, number of examples, etc. Past studies (arguably without the same controls for similarity) have shown evidence of better learning for linearly separable categories with more clearly differentiated categories, more than two categories, more examples per category, non-binary dimensions, instructions that highlight integrated feature encodings, and tasks other than classification learning (Blair & Homa, 2001; Smith et al., 1997; Wattenmaker, Dewey, Murphy, & Medin, 1986; Yamauchi, Love, & Markman, 2002). We agree with the broad perspective and join the many voices calling for richer stimuli, categories, and tasks (see Kurtz, 2015). Useful future work should look more closely at the factors contributing to cases in which NLS categories are or are not learned well, as well as the prevalence of linear separability in naturally occurring categories (see Ruts, Storms, & Hampton, 2004). Nonetheless, the use of small, carefully controlled categories like the ones in the current study conforms to a scientific approach that has taught researchers a great deal about the psychology of human category learning – and formal models of categorization that represent our best hope for generalizable theory should be evaluated and improved upon as a result of such data.

## Model simulations

As discussed above, the original Medin and Schwanenflugel (1981) report was highly influential in guiding theoretical development in the study of human categorization – concrete evidence against a core prediction of prototype models

<sup>3</sup> It would follow that there are other pure memorizers learning other items (prototypes, intermediate) first. Unfortunately, these learners would be difficult to identify from the data because their performance profile would happen to match closely that of other strategies discussed.

rendered them essentially a non-contender in the playing field of formal modeling of the traditional artificial classification learning paradigm. We had two main goals in the cognitive simulations: (1) to test contemporary implementations of the exemplar and prototype views relative to the revised behavioral pattern showing a robust NLS advantage during training; and (2) to test similarity-based models that take a more sophisticated approach to learning abstraction-based solutions.

ALCOVE (Kruschke, 1992) is an exemplar-based adaptive network model that classifies items according to attentionally-weighted similarity to known exemplars. ALCOVE is a process model that builds on the advances of Nosofsky (1986) in generalizing the context model in terms of stimulus generalization theory and adds two design properties: error-driven learning of attentional weights and the inclusion of dynamic, adaptive association weights between exemplars and classes. ALCOVE takes a stimulus as input (as experimenter-defined if perfectly clear or based on multidimensional scaling to estimate psychological representations otherwise) and recodes it according to its similarity (using inverse exponential distance) to a reference point node for each item in the training set. The similarity is mediated by dimensional selective attention that shrinks or stretches the geometric space according to dimensional diagnosticity – these attention weights are optimized through error-driven learning. A set of traditional neural network weights fully connect the exemplar nodes to the output layer consisting of class nodes. Error-driven learning is used to set an associative strength between each exemplar and each category. This implements the core design principle that an item is categorized as an A to the extent it is highly similar (under a set of dimensional weights) to known As and not highly similar to known Bs. A specificity parameter determines how sharply the reference points fall off in their receptive fields (i.e., to what extent they have overlapping regions). The final step is a response mapping function to turn the activation of the class nodes into a probabilistic response.

Rather than employing a minimal implementation of the prototype view, we wanted to give it all the potential advantages of ALCOVE with the critical difference of using prototype reference points instead of exemplars. This means including inverse exponential distance metric (Shepard, 1957, 1987) with a free parameter for sensitivity/specificity, dimensional selective attention, and a Luce choice rule with a free parameter for response mapping (see Kruschke, 1992 for details). This “souped-up” prototype model (PROTO-ALCOVE) differs from ALCOVE only in that similarity is evaluated relative to a prototype for each category, rather than to the exemplars themselves – an approach that has been taken in the past (e.g., Nosofsky & Stanton, 2005). We tested two versions: prototype as the mean value across category members for each dimension vs using the mode value (since the features are binary). All models tested were conveniently equivalent in having four free parameters. In the case of ALCOVE and

PROTO-ALCOVE, these were: the specificity of exemplar/prototype generalization ( $c$ ), the association learning rate ( $\lambda_w$ ), the attention learning rate ( $\lambda_a$ ), and response mapping ( $\phi$ ).

In accord with our goal of evaluating abstraction-based models that do more than substitute category central tendency for the individual exemplars, we tested the Supervised and Unsupervised STRatified Adaptive Incremental Network (SUSTAIN: Love et al., 2004) model – an adaptive clustering approach that expands its architecture in response to unexpected events by adding clusters that can represent individual examples or groups of highly similar examples. This model occupies a “middle-ground” within the reference point framework between the extremes of storing every example and storing a single summary prototype per category. SUSTAIN is much like ALCOVE in its use of an input layer that projects to a layer of reference point nodes that in turn project to a set of class nodes via feedforward layer weights. The differences (full details are beyond the scope of the present discussion) include: a dynamic process for building clusters by taking new stimuli that are well handled by an existing cluster and adjusting that cluster to the new centroid or dealing with “surprising” items (that are not well handled) by creating a new reference point node centered on that item’s dimension values; a mechanism for adjusting dimensional attention that is not error-driven; a competitive element to mediate the impact across clusters; and an ability to engage in feature prediction (that is not used in the standard classification learning mode). The four free parameters in SUSTAIN are: attentional focus ( $r$ ), cluster competition ( $\beta$ ), learning rate ( $\eta$ ), and response mapping ( $d$ ). We note that attentional learning is data-driven rather than error-driven in SUSTAIN (see Love et al., 2004 for further details).

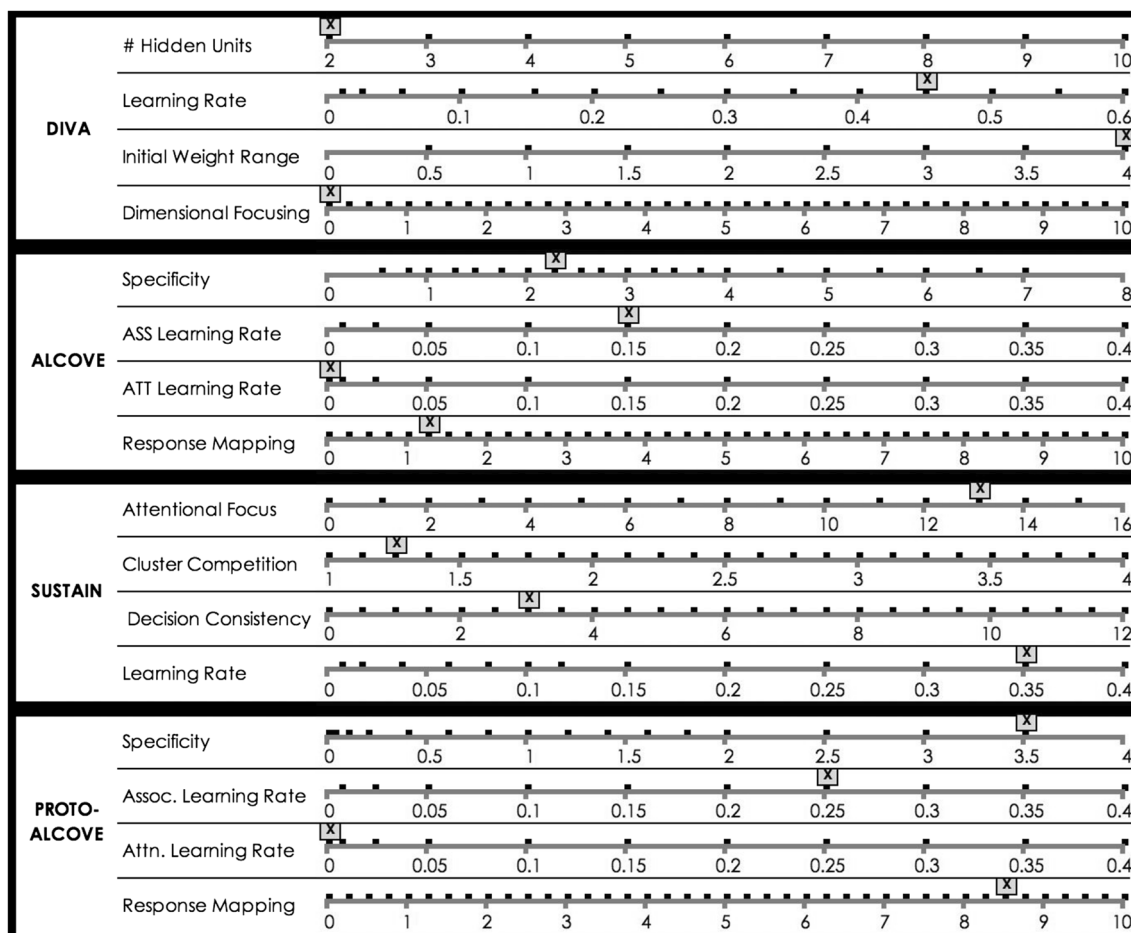
Finally, we conducted simulations with the DIVERgent Autoencoder (DIVA: Kurtz, 2007) model. Unlike the approaches described above, DIVA is similarity based without invoking the stimulus generalization framework of inverse exponential distance to localist reference points (see Conaway & Kurtz, 2016; Kurtz, 2015). DIVA is a connectionist network that interprets the problem of learning categories in terms of within-class feature prediction: following a more generative approach (Ng & Jordan, 2002), the network uses representation learning via backpropagation of error to project the training domain into a constructed feature space (a standard connectionist hidden layer) so that an auto-associative channel for each category can reconstruct the original feature values (for more detailed discussion, see Conaway & Kurtz, 2017; Kurtz, 2007, 2015). Therefore the critical differences from ALCOVE are: (1) recoding by representation learning rather than similarity to reference points; (2) predicting a construal (set of feature values) of the input with respect to the learned generative model of each category and making classification decisions based on how well the features of the stimulus are preserved through the process of recoding and

reconstruction along each category channel; and (3) applying dimensional focusing as part of the response rule by prioritizing dimensions that make diverse predictions across the categories. Following standard practices with DIVA, we used a sigmoid activation rule (in accord with binary dimension stimuli), encoded the stimuli using +/-1 input values and 0/1 targets, and fitted four free parameters: the number of hidden units, the range of the initial random weights, the learning rate, and the dimensional focusing parameter ( $\beta$ ).

**Qualitative model evaluation** The parameters of each model were fitted to the aggregate human learning data using a grid-search procedure – measuring classification accuracy for the two category structures across a wide range of parameter settings (see Fig. 7 for ranges of search parameters). First, we considered the best-fitting parameterizations for each model (see Fig. 4). With regard to the classic exemplar versus prototype comparison: ALCOVE captured the NLS advantage while PROTO-ALCOVE did not. The two more sophisticated abstractive models, DIVA and SUSTAIN also captured a clear NLS advantage. DIVA works by learning a set of within-category feature prediction tasks. Given the

LS category (100, 010, 001), DIVA seeks a recoding of each item in the latent space such that for ease of predicting the first feature items “010” and “001” (because they match on the first feature) are proximally located; for ease in predicting the second feature, items “100” and “001” are proximal, and for ease in predicting the third feature, items “100” and “010” are proximal. Note that each of these critical pairs is maximally dissimilar (mismatching on two dimensions). In the NLS case (001, 010, 110), there is less dissimilarity among within-category item pairs that need to be recoded proximally – and therefore easier learning. SUSTAIN’s tendency is to learn an LS category (001, 010, 100) by creating a cluster of two items based on selective attention to a single matching dimension value and an NLS category (001, 010, 110) by creating a cluster of two items based on selective attention to two matching dimensions (?10). Note that in the LS structure, the model must cope with two members of the other category matching the cluster; in the NLS structure, there are no members of the other category that match the cluster.

An advantage of the grid search approach is that we are able to systematically evaluate each model’s range of



**Fig. 7** Parameter value search space for the four model simulations. The boxes along each number line show the range and granularity of the grid search for each model parameter. The best fit for each parameter is indicated with an X

predictions across parameterizations in addition to finding a best fitting parameterization. We computed the strength of the NLS advantage under each parameterization using a difference score metric (difference = LS-NLS). Positive difference scores indicate that a given parameterization produced greater accuracy on the LS problem, whereas negative difference scores indicate greater accuracy on the NLS problem. By plotting the distribution of differences scores across all parameterizations, we can visualize the possible qualitative results that each model is capable of explaining. These data, shown in Fig. 8, reveal a clear pattern. While DIVA, ALCOVE, and SUSTAIN most commonly predict greater accuracy on the NLS problem, the PROTO-ALCOVE model most commonly predicts no difference. The high frequency of PROTO-ALCOVE parameterizations that produce equal LS and NLS learning reflects the fact that the model usually fails to achieve above-chance accuracy in either condition: the prototypes, localized as the average (mean or mode) of each category's examples, are somewhat distant from the exemplars themselves. As a result, the prototypes do not generalize strongly onto their examples unless they are given a wide generalization range (using low specificity,  $c$ , values) which produces poor learning (and therefore a poor qualitative fit). To be clear, the failure of the PROTO-ALCOVE model to learn the LS structure under many of the tested parameterizations is not a shortcoming of the model (any model will fail catastrophically

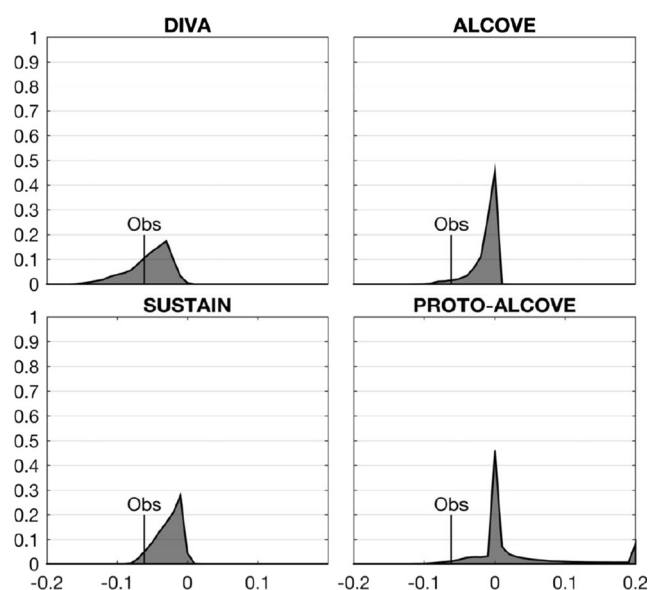
under parameter settings that are discordant with its design principles); the important take-away is that with appropriate parameter settings, PROTO-ALCOVE embodies the prototype view fully in terms of its ability to readily acquire the LS structure and its inability to acquire the NLS structure (as shown in Fig. 4).

As pointed out by an anonymous reviewer, it would be informative (albeit impossible) to have a density profile for the human behavioral data, so we must rely on the high-power sample as an estimate of the phenomenon being fit. Finally, with regard to the concern sometimes raised about formal models potentially being overpowered, an important observation is that none of the successful models predict the reverse finding (i.e., an LS advantage) under any parameterization.

## Conclusions

We tested the category structures from Medin and Schwanenflugel (1981, Exp. 4) using standardized materials, more training, and more power – and found that participants learning the NLS classification made significantly fewer errors than those learning the LS classification. To our knowledge, a robust NLS advantage has never been found using category structures fully matched in within- and between- category additive similarity. For NLS learners, items possessing the most common features were easier to learn and were mastered most successfully, a majority of participants found the exception item harder to learn and less typical than other items. A subset of learners found the exception items easier to learn and showed no clear typicality pattern, a finding that is problematic for current category learning accounts. These results provide clear evidence against a linear separability constraint given the materials and category structures used in this study.

The modeling results demonstrate that an NLS advantage is fully consistent with exemplar-based category learning and directly contradicts a prototype approach (even implemented with maximal advantages). Further, we see that more sophisticated abstractive-based similarity accounts predict the NLS advantage. In closely related research, we are pursuing a comprehensive investigation of the power of competing formal models to account for the detailed patterns of behavioral data in this report – including quantitative modeling of individual/subgroup and item-level differences in performance. For present purposes, the critical point to convey from a modeling perspective is that a prototype-based account is unable to predict the NLS advantage under any parameterization, while similarity-based models that instantiate a range of alternative theoretical positions (ALCOVE, SUSTAIN, DIVA) readily predict the NLS advantage.



**Fig. 8** Density profiles representing proportion of fits predicting various learning outcomes across the range of parameterizations tested for each model. A zero indicates the prediction that linearly separable (LS) and non-linearly separable (NLS) categories would be learned equally well; negative values indicate an NLS learning advantage. *Obs* indicates the observed effect in the behavioral data. Note that the spike at the zero-difference level in the density profile for the PROTO-ALCOVE model reflects parameterizations that lead to no learning of either category structure



**Open practices statement** Behavioral data for this experiment is available on the Open Science Foundation website at <https://osf.io/x3gqe/>. The experiment was not preregistered.

## References

- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409-429.
- Blair, M. & Homa, D. (2001). Expanding the search for a linear separability constraint on category learning. *Memory & Cognition*, *29*, 1153-1164.
- Blair, M. & Homa, D. (2003). As easy to memorize as they are to classify: The 5-4 categories and the category advantage. *Memory & Cognition*, *31*, 1293-1301.
- Conaway, N. B. & Kurtz, K. J. (2016). Generalization of within-category feature correlations. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the Thirty-Eighth Conference of the Cognitive Science Society* (pp. 2375-2380). Austin, TX: Cognitive Science Society.
- Conaway, N. B. & Kurtz, K. J. (2017). Similar to the category, but not the exemplars: A study of generalization. *Psychonomic Bulletin & Review*, *24*, 1312-1323.
- Homa, D., Rhoads, D., & Chambliss, D. (1979). Evolution of conceptual structure. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 11-23.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22-44.
- Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3-36.
- Kurtz, K. J. (2007). The divergent autocoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*, 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*, 552-572.
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309-332.
- Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *TRENDS in Cognitive Sciences*, *6*, 255-260.
- Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *8*, 37-50.
- Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.
- Medin, D. L. & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 355-368.
- Ng A. & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naïve Bayes. *Advances in Neural Information Processing Systems*, *14*, 841-848.
- Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *14*, 54-65.
- Nosofsky R. M. & Stanton, R. D. (2005). Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 608-629.
- Nosofsky, R. N., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53-79.
- Posner, M. I. & Keele, S. (1968). On the genesis of abstract ideas. *Journal of experimental Psychology*, *77*, 491-504.
- Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382-407.
- Rosch, E. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573-605.
- Ruts, W. Storms, G., & Hampton, J. (2004). Linear separability in superordinate natural language concepts. *Memory & Cognition*, *32*, 83-95.
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.
- Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325-345.
- Shepard, R. N., Hovland, C. I. & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*, 1-42.
- 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*, 1411-1436.
- Smith, J. D., Murray, M. J., & Minda, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 659-680.
- Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology*, *18*, 158-194.
- 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*, 585-593.

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