

Hold-out strategy for selecting learning models: application to categorization subjected to presentation orders

Giulia Mezzadri^{1,*}, Thomas Laloë¹, Fabien Mathy², and Patricia Reynaud-Bouret¹

¹Université Côte d'Azur, Laboratoire J.A. Dieudonné UMR CNRS 7351, Nice, France

²Université Côte d'Azur, Bases, Corpus, Langage UMR CNRS 7320, Nice, France

Abstract


In this article, we develop a new general inference method for selecting learning models. The method relies upon a specific hold-out cross-validation, which takes into account the dependency within the data. This allows us to retrieve the model that best fits the learning strategy of a single individual. The novelty of our approach lies on the choice of the testing set, both in the experimental design and in the data analysis. This individual approach is then applied to two category learning models (ALCOVE and Component-cue) on data-sets manipulating presentation order, after verification of the reliability of our method. We found that both models performed equally well during transfer, but Component-cue best fits the majority of participants during learning. To further analyze these models, we also investigated a potential relation between the underlying mechanisms of the models and the actual types of presentation order assigned to participants.

Keywords: model selection, learning models, statistical inference, hold-out cross-validation, category learning, Component-cue, ALCOVE, rule-based order versus similarity-based order

Introduction

Computational models are now common in many domains of cognitive science, for instance to study memory (Lemaire & Portrat, 2018; Oberauer & Kliegl, 2006), decision making (Ariofovic & Ledyard, 2011; Novikov et al., 2018; Roth & Erev, 1995), attention (Borji & Itti, 2013; Malem-Shinitski et al., 2020), and categorization (Carvalho & Goldstone, 2022; Kruschke, 1992; Love et al., 2004). Because formal models abound in cognition, methods have been developed to offer rigorous common grounds to evaluate their performance (Myung, 2000; Myung & Pitt, 1998; Pitt et al., 2002). The purpose of our study is to promote the use of a general method to fit formal learning models to experimental data, with a particular focus on models of category learning.

* Corresponding author.

E-mail address: gm3026@columbia.edu (G. Mezzadri)  <https://orcid.org/0000-0001-6453-9070>.

In categorization, a large variety of practices exist to fit models to data. Some studies have used the same set of observations to both estimate the parameters of the models and compute their predictions (Nosofsky et al., 2017, 2018; Sanders & Nosofsky, 2020), running the risk of over-fitting the data (Cawley & Talbot, 2010). Other studies have relied on the use of computer simulations with the aim of either estimating the parameters of the models or determining their predictions (Carvalho & Goldstone, 2022; Nosofsky et al., 1994, 2017). In these studies, the estimated parameters and the overall predictions of the models were obtained by averaging the best-fitting parameters and the classification predictions found in the simulations. Finally, a wide array of criteria have been used to estimate fit to the data (Carvalho & Goldstone, 2022; Nosofsky et al., 1992, 1994, 2018). Some examples are the Sum of Squared Deviations (SSD), the Weighted Sum of Squared Deviations (WSSD), the likelihood (either trial-by-trial, or block-by-block, or epoch-by-epoch), the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC).

In addition to this heterogeneity of practices, in contexts when observations are not independent (such as learning) classical statistical criteria are not reliable. While these criteria have statistical guarantees when applied to independent and identically distributed (i.i.d.) data, such guarantees are no longer available with dependent data. For instance, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) offer theoretical guarantees when the number of observations tends to infinity and when observations are i.i.d. (Akaike, 1998; Claeskens & Hjort, 2007; Konishi & Kitagawa, 2008; Schwarz, 1978). Since participants learn during categorization tasks and since their learning process ends in finite time, participants' observations are dependent on one another and their number is limited.

In this article, we propose a new statistical inference method for model selection that can be applied in contexts involving learning. Our method falls under the category of cross-validation methods, that are generally more flexible than classical statistical criteria (Allen, 1974; Stone, 1974). Here, we propose the simplest kind of cross-validation: the hold-out which consists in separating observations in two sets, one for parameter estimation and one for model testing. However, training and testing sets are not randomly selected as in usual cross-validation methods. Indeed, data of a single individual acquired from learning tasks cannot be thought independent and this dependency structure within the data need to be precisely taken into account. Therefore, models are tested either on the transfer phase (when the object of interest is performance during transfer) or on unsupervised blocks of the learning phase (when the object of interest is learning progression). Learning refers to the stage in which categories are formed, while transfer refers to the stage in which individuals' knowledge is tested upon presentation of new stimuli. However, the learning phase is generally exclusively composed of supervised blocks (Carvalho & Goldstone, 2014; Mathy & Feldman, 2009). Therefore, with the aim of applying our inference method to the learning phase alone, we specifically designed and conducted an experiment including unsupervised blocks. This specific experimental design is completely original in learning experiments. Up to our knowledge, this is the first design which, combined with the adequate statistical analysis, allows us to determine which model best fits a learning phase.

Although our method is flexible enough to be applied to all kinds of learning models, here it is applied to compare two models of category learning. The two models on which our investigation is focused are Gluck and Bower's Component-cue (Gluck & Bower, 1988) and Kruschke's ALCOVE (Kruschke, 1992). Both models have the ability to evolve over time, accounting for both category learning and transfer. The selection of these two models was motivated by the fact that, although their mathematical structure is similar, they implement different learning strategies. In-

deed, both models are based on artificial neural networks (Dreyfus, 1990; Rosenblatt, 1958); however, they implement either a complex rule-based strategy (Component-cue) or a similarity-based strategy (ALCOVE) (Högden et al., 2019). A complex rule-based strategy refers to the process with which participants classify new items on the basis of complex previously acquired rules, whereas a similarity-based strategy refers to the process with which participants classify new items on the basis of their similarity to stored exemplars or prototypes. Analyzing models with a similar mathematical architecture allowed us to focus on the psychological mechanisms implemented into these respective models. Our goal was to determine whether a complex rule-based or a similarity-based strategy best fits our data-sets, after showing statistical guarantees of our inference method through numerical simulations. Also, these numerical simulations supported the use of an individual approach, in which each participant is solely fit.

To further analyze these models, we made use of two types of presentation order involving a variation of stimulus ordering within a category (Bower et al., 1969; Medin & Bettger, 1994). The rule-based order is designed to facilitate a rule-abstraction process ordering stimuli following a “principal rule plus exceptions” structure, whereas the similarity-based order is designed to maximize the similarity between consecutive stimuli (Elio & Anderson, 1981, 1984; Mathy & Feldman, 2009, 2016). The rationale is that a model should perform better when stimuli are presented following a presentation order inspired by the mechanisms at play in the model. For instance, a model integrating a rule-based or a complex rule-based strategy should benefit from a presentation in which stimuli obeying the principal rule are presented before the exceptions. Inversely, a model integrating a similarity-based strategy should benefit from a presentation that maximizes the similarity between contiguous examples. In both cases, the external context that best suits the internal mechanism of a model should facilitate the extraction of the categories. Therefore, our hypothesis is that Component-cue should best fit participants in the rule-based order, while ALCOVE should best fit participants in the similarity-based order.

To summarize, the objective of this article is three-fold: *i*) to present a general method to guide the evaluation and selection of learning models (Section [Statistical Inference Method](#)), *ii*) to apply this method on designed data-sets for comparing two category learning models (Component-cue and ALCOVE) that implement different learning strategies (Section [Model-fitting results](#)), and *iii*) to investigate whether the learning strategies at play in the models are related to the order in which stimuli are presented, when the chosen order is inspired by these strategies (Section [Relation between models and within-category orders](#)). Numerical simulations validating our inference method are given in Section [Numerical simulations](#), following the description of the models and data-sets (Section [Overview of two models of category learning](#) and [Overview of the data-sets](#)).

Statistical Inference Method

Here, we first describe how the parameters of the models are estimated. Then, we present the statistical inference method used to determine which model best accounts for category learning and transfer. Numerical simulations assessing the accuracy of the estimates of the chosen models, as well as the reliability of the method are given in the section Results.

Parameter Estimation

The parameters of the models were estimated using the Maximum Likelihood Estimation (MLE) (Aldrich, 1997):

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \{-\log \mathcal{L}_M(D; \theta)\},$$

where M denotes the model, \mathcal{L}_M its likelihood, and D the data-set used for the estimation. The MLE was performed using the gradient descent algorithm. To avoid local minima, the gradient descent algorithm was run 10 times, taking each time different initial conditions.

Model selection

Models were fit to our data-sets using the hold-out cross-validation method, which consists in training the models on a subset of the data and testing them on the remaining subset. As discussed in the introduction, the use of a cross-validation technique was preferred to classical statistical criteria because of its flexibility and ability to be applied in contexts involving learning. Because it would have been too intricate to apply convoluted cross-validation techniques on non-i.i.d. data, the simplest kind of cross-validation (hold-out) was adopted. When cross-validation techniques are applied to i.i.d. data, training and testing sets are completely exchangeable. However, as mentioned above our observations during learning are dependent on one another because of feedback. This dependency within the data makes it extremely difficult to train models on observations that occur after the observations on which models are tested (this would require a very complex “expectation-maximisation” phase that is out of the scope of the present paper). Potential solutions are to either train models on observations that occur before the observations on which models are tested, or to test models on observations with no feedback and on which a “frozen” model that does not evolve is used. The latter is used here, whereas the former has been used in spatial learning tasks for non-human animals (Moongathottathil-James et al., 2021). In both cases, the method is a particular case of hold-out where the testing set has to be intentionally well-chosen.

The predictions of the models were evaluated with either the Sum of Squared Deviations (SSD) or the likelihood. The SSD is given by the sum of the squared difference between the prediction of the model and the participants’ response across the testing set:

$$E_{\text{SSD}}(M) = \sum_{x^{(t)} \in D_T} \left(\mathbb{P}_M^{\hat{\theta}}(A | x^{(t)}, \mathcal{H}_{t-1}) - z^{(t)} \right)^2,$$

where M denotes the model; $\mathbb{P}_M^{\hat{\theta}}(A | x^{(t)}, \mathcal{H}_{t-1})$ is the prediction of the model for the stimulus $x^{(t)}$, given the sequences of stimuli and feedback \mathcal{H}_{i-1} until time $t-1$; $z^{(t)}$ is the response given by the participant for the classification of the stimulus $x^{(t)}$; and D_T is the testing set. The parameter $\hat{\theta}$ was estimated on the training set.

The evaluation of the model using the likelihood is given by:

$$E_{\mathcal{L}}(M) = -\log \mathcal{L}_M(D_T; \hat{\theta}),$$

where M denotes the model; \mathcal{L}_M its likelihood; $\hat{\theta}$ the estimated parameter on the training set; and D_T the testing set. The model that best fit our data-sets \hat{M} is the model with the lowest evaluation

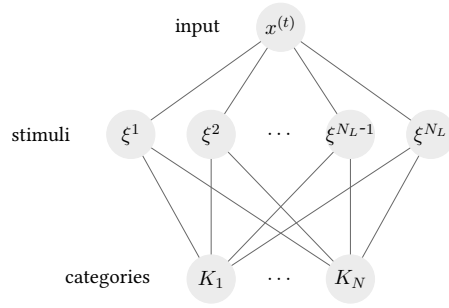


Figure 1

Artificial neural network structure of ALCOVE.

with either the SSD or likelihood criteria ($*$ = SSD or \mathcal{L}):

$$\hat{M} \in \arg \min_M \{E_*(M)\}.$$

Although we advise the use of the likelihood criterion when the parameter estimation is performed using MLE, we additionally considered SSD in order to allow a continuity with previous studies in categorization. Indeed, the use of SSD in psychology is still popular (Carvalho & Goldstone, 2022; Nosofsky et al., 1992, 1994, 2018; Palmeri, 1999).

Overview of two models of category learning

Here, we present the category learning models that we compared using the statistical inference method described above. As mentioned in the introduction, ALCOVE and Component-cue have a similar mathematical structure, but implement different learning strategies. ALCOVE learns the category membership of the training stimuli and classifies new items on the basis of their similarity to these acquired stimuli. Conversely, Component-cue learns the combination of features that are a good predictor of the category membership of the training stimuli and classifies new items on the basis of these diagnostic combination of features.

ALCOVE

Attention Learning COVERing map model (ALCOVE) (Kruschke, 1992) is an artificial neural network composed of three layers of nodes: *i*) a single input node receiving the stimuli, *ii*) a layer of intermediate nodes coding for the learning stimuli, and *iii*) a layer of output nodes coding for the categories in which stimuli can be classified (see Figure 1). The intermediate nodes are linked to the output nodes through association weights, whose evolution allows the model to learn. When a stimulus $x^{(t)}$ reaches the input node, the intermediate nodes ξ^j (for $j = 1, \dots, N_L$) are activated by the quantity:

$$a_j^{(t)} = S(x^{(t)}, \xi^j),$$

where N_L is the number of learning stimuli. The term $S(x^{(t)}, \xi^j)$ denotes the similarity between stimuli $x^{(t)}$ and ξ^j , and it is computed as an exponentially decaying function of the distance between the two stimuli:

$$S(x^{(t)}, \xi^j) = e^{-c \cdot d(x^{(t)}, \xi^j)^p},$$

where $d(x^{(t)}, \xi^j)$ is the distance between stimuli $x^{(t)}$ and ξ^j , p a positive constant, and c a freely estimated sensitive parameter ($c \geq 0$). The distance between stimuli $x^{(t)}$ and ξ^j is computed as follows:

$$d(x^{(t)}, \xi^j) = \left[\sum_{i=1}^{\mathfrak{N}} \omega_i \cdot |x_i^{(t)} - \xi_i^j|^r \right]^{\frac{1}{r}},$$

where \mathfrak{N} is the dimension of the psychological space in which stimuli are embedded, ω_i the attention allocated to dimension i ($\omega_i \geq 0$ and $\sum_{i=1}^{\mathfrak{N}} \omega_i = 1$), r a positive constant; and $x_i^{(t)}$ and ξ_i^j the feature values of stimuli $x^{(t)}$ and ξ^j on dimension i . The values of p and r are determined on the basis of the nature of the stimuli. In our case, p and r are set equal to 1 (see Section [Overview of the data-sets](#)).

All these quantities $a_j^{(t)}$ (for $j = 1, \dots, N_L$) are weighted and summed to form outputs. The output node associated with category K is activated by the quantity:

$$O_K^{(t)} = \sum_{j=1}^{N_L} a_j^{(t)} \cdot w_{j,K}^{(t)},$$

where $w_{j,K}^{(t)}$ is the association weight linking intermediate node ξ_j to output node K , at the arrival of the t -th stimulus. The outputs were constrained to vary between -1 and 1. The probability of classifying a stimulus into a given category is computed as a function of the outputs. Two formulas have been used in the literature: an exponential formula ([Kruschke, 1992](#)) and a linear formula ([Nosofsky et al., 1992, 1994; Palmeri, 1999](#)). According to the exponential formula, the probability of classifying the t -th stimulus $x^{(t)}$ as belonging to a given category A (knowing the sequence of stimuli and feedback \mathcal{H}_{t-1} until time $t - 1$) is given by:

$$\mathbb{P}(A | x^{(t)}, \mathcal{H}_{t-1}) = \frac{e^{\phi O_A^{(t)}}}{\sum_{K \in \mathcal{K}} e^{\phi O_K^{(t)}}}, \quad (1)$$

where ϕ is a freely estimated positive parameter and \mathcal{K} the set of all categories. According to the linear version, the probability of classifying the t -th stimulus $x^{(t)}$ as belonging to a given category A (knowing the sequence of stimuli and feedback \mathcal{H}_{t-1} until time $t - 1$) is given by:

$$\mathbb{P}(A | x^{(t)}, \mathcal{H}_{t-1}) = \frac{O_A^{(t)} + b}{\sum_{K \in \mathcal{K}} (O_K^{(t)} + b)} \quad (2)$$

where b is a category bias parameter ($b \geq 1$) and \mathcal{K} the set of all categories. The exponential version is denoted by the letter E (i.e., ALCOVE^E), while the linear version by the letter L (i.e., ALCOVE^L). In our study, both formulas were considered.

Once the classification probability are computed, the association weights $w_{j,K}^{(t)}$ and attention weights $\omega_i^{(t+1)}$ are updated in order to minimize the difference between feedback and outputs of the model. More specifically, the error of the model for the t -th stimulus is computed as follows:

$$E^{(t)} = \sum_{K \in \mathcal{K}} (\mathcal{T}_K^{(t)} - O_K^{(t)})^2,$$

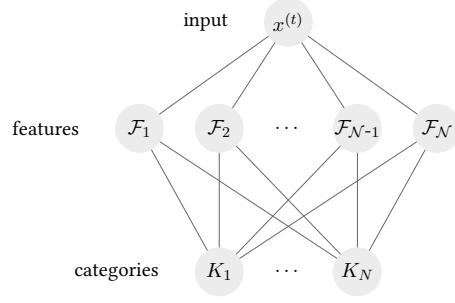


Figure 2

Artificial neural network structure of Component-cue.

where

$$\mathcal{T}_K^{(t)} = \begin{cases} 1 & \text{if } x^{(t)} \in K \\ -1 & \text{otherwise} \end{cases}$$

The association weights are updated to decrease the error of the model as follows:

$$\begin{aligned} w_{j,K}^{(t+1)} &= w_{j,K}^{(t)} - \lambda_w \cdot \frac{\partial E^{(t)}}{\partial w_{j,K}^{(t)}} \\ &= w_{j,K}^{(t)} + \lambda_w \cdot a_j^{(t)} \cdot \left(\mathcal{T}_K^{(t)} - O_K^{(t)} \right), \end{aligned} \quad (3)$$

where λ_w is a freely estimated learning rate parameter ($\lambda_w \geq 0$). The attention weights are equally updated to decrease the error of the model. Their updating is given by the following rule:

$$\begin{aligned} \omega_i^{(t+1)} &= \omega_i^{(t)} - \lambda_\omega \cdot \frac{\partial E^{(t)}}{\partial \omega_i^{(t)}} \\ &= \omega_i^{(t)} - \lambda_\omega \cdot \sum_{K \in \mathcal{K}} \sum_{j=1}^{N_L} \left(a_j^{(t)} \cdot w_{j,K}^{(t)} \cdot c|x_i^{(t)} - \xi_i^j| \cdot \left(\mathcal{T}_K^{(t)} - O_K^{(t)} \right) \right), \end{aligned}$$

where λ_ω is a freely estimated learning rate parameter ($\lambda_\omega \geq 0$). The association and attention weights are initiated at 0. ALCOVE can also be applied to reproduce performance during transfer. In this scenario, since feedback is not provided during transfer, the weights are no longer updated and a “frozen” model is considered. This is also true on unsupervised blocks of the learning phase where participants’ classification is monitored without feedback.

Component-cue

Component-cue (Gluck & Bower, 1988) is an artificial neural network, composed of three layers of nodes: *i*) a single input node receiving the stimuli, *ii*) a layer of intermediate nodes coding for the features of the stimuli, and *iii*) a layer of output nodes coding for the categories in which stimuli can be classified (see Figure 2). As in ALCOVE, the intermediate nodes are linked to the output nodes through association weights, whose evolution allows the model to learn. When a

stimulus $x^{(t)}$ reaches the input node, the intermediate nodes \mathcal{F}_j ($j = 1, \dots, \mathcal{N}$) are activated as follows:

$$a_j^{(t)} = \begin{cases} 1 & \text{if } x^{(t)} \text{ has } \mathcal{F}_j \\ 0 & \text{otherwise.} \end{cases}$$

All of these quantities $a_j^{(t)}$ (for $j = 1, \dots, \mathcal{N}$) are weighted and summed to form outputs. The output node associated with category K is activated by the quantity:

$$O_K^{(t)} = \sum_{j=1}^{\mathcal{N}} a_j^{(t)} \cdot w_{j,K}^{(t)},$$

where $w_{j,K}^{(t)}$ is the association weight linking intermediate node \mathcal{F}_j to output node K . Again, the outputs were constrained to vary between -1 and 1. Similarly to ALCOVE, the classification probabilities are computed as in Equation (1) (if the exponential formula is considered) or as in Equation (2) (if the linear formula is considered). The same notations as before are used to denote the two versions. Again, once the classification probability are computed, the association weights are updated in order to minimize the error of the model. The association weights of Component-cue are updated as in Equation (3) and their initialization is set at 0.

Overview of the data-sets

Models were compared based on two separate data-sets. The first data-set corresponds to the results of an experiment conducted by (Mathy & Feldman, 2016), which was designed to assess the effects of within-category orders on category transfer. The second data-set corresponds to the results of an experiment conducted by (Mezzadri et al., 2022). Although this second data-set has already been used to test a model of category transfer, it was specifically designed for the application of our method to both the learning phase alone and the totality of the experiment. As mentioned in the introduction, the learning phase of a categorization task is generally supervised. Supervision implies dependency of the observations, which heightens the complexity of the application of cross-validation techniques. The introduction of unsupervised blocks within the learning phase allowed us to apply our method, without increasing its complexity. Although they are not novel experiments, the procedure of both experiments is briefly recalled.

Data-set 1

Participants ($N = 44$) were instructed to learn a 4-feature category structure (see Figure 3) based on either a rule-based presentation order or a similarity-based presentation order. This structure, called 5-4 category set (Medin & Schaffer, 1978), allowed to study how participants categorize 7 novel stimuli during a transfer phase, after learning $5+4 = 9$ stimuli (5 items belonged to category A and 4 items to category B). Participants were instructed to press one of two response keys corresponding to the categories. A feedback indicating the correctness of their responses was provided, except in the transfer phase.

Stimuli. Stimuli varied along four Boolean dimensions (shape, color, size, and filling pattern). The options for each dimension were: square or circle for shape; blue or gray for color; small or big for size; and plain or striped for filling pattern. The combination of these features formed $2^4 = 16$ items (Figure 3, on the bottom). Each dimension was instantiated by the same

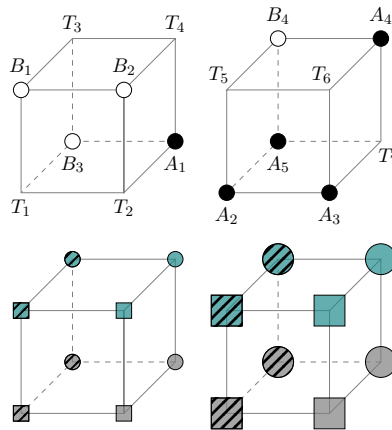


Figure 3

Categories and stimulus items of Data-set 1. The items varied along four Boolean dimensions (shape, color, size and filling pattern) represented here in a Hasse Diagram forming a hypercube. At the top, the structure of the 5-4 category set. The examples of category A are indicated by black dots, those of category B by white dots, and transfer items are represented by empty vertices. At the bottom, illustration of the items of Data-set 1.

physical feature for all participants. As can be seen in Figure 3 (on the bottom), color differentiated the objects at the top of the hypercube from those at the bottom, shape differentiated the objects at the front of the hypercube from those at the back, size distinguished the objects in the left cube from those in the right cube, and filling pattern differentiated the right and left objects within the cubes.

Phases. The experiment was composed of a supervised learning phase (in which feedback was provided at each trial), followed by an unsupervised transfer phase (in which no feedback was provided). Participants had to correctly classify stimuli in four consecutive blocks of 9 stimuli to complete the learning phase. Once participants met this learning criterion, a transfer phase was initiated. The transfer phase was composed of 5 blocks of 16 stimuli (the 9 learning items plus 7 transfer items).

Ordering of stimuli. During learning, training blocks were alternated with random blocks. Training blocks were used to manipulate order, while random blocks were used to monitor learning. In training blocks, categories were blocked (i.e., *AAAABBBB* or *BBBBAAAA*) and the order of the stimuli within a category was manipulated following either a rule-based or a similarity-based order. Half of the participants were randomly assigned to the similarity-based condition. In the rule-based order, stimuli obeying the main rule were presented strictly before the exceptions to the rule. The principal rule was determined by the color (all gray items are members of category A and all blue items are members of category B), while the exceptions were the small gray hatched circle and the big blue plain circle. In the similarity-based order, members within a category were presented in a way that maximized the similarity between consecutive stimuli. The first stimulus was randomly selected and subsequent stimuli were (randomly) selected among those that were the most similar to the immediately previous item. Similarity between two items was computed by counting the number of common features that they shared and ties were solved

randomly. For further details we refer the reader to (Mathy & Feldman, 2016).

Data-set 2

As in the previous data-set, participants ($N = 130$) were instructed to learn a single 5-4 category set based on different types of order. Although the categories and stimuli were similar to that of Data-set 1, Data-set 2 extended the manipulation of presentation orders. In addition to the within-category manipulations, both between-category and across-blocks conditions were manipulated. These variations were introduced to avoid picking a condition which could favor one of the two condition of our main factor (rule-based or similarity-based). To summarize, Data-set 2 extends Data-set 1 on three levels: *i*) the introduction of unsupervised blocks during learning, allowing the application of our method to the learning phase alone, *ii*) the larger variety of order manipulations, and *iii*) the higher number of participants.

Stimuli. Stimuli were the same as in Data-set 1. However, dimensions were instantiated by different features. Indeed, color distinguished the objects at the front of the hypercube from those at the back, shape distinguished the objects in the left cube from those in the right cube, size distinguished the right and left objects within the cubes, and filling pattern distinguished the objects at the top of the hypercube from those at the bottom.

Phases. As in Data-set 1, a learning phase was followed by a transfer phase. However, Data-set 2 made use of two blocks of supervised learning (in which the order of the stimuli was manipulated and feedback was provided), followed by one block of unsupervised learning (in which stimuli were randomly presented with no feedback). This pattern was repeated until the end of the learning phase. Participants had to correctly classify stimuli in three unsupervised blocks of 9 stimuli (not necessarily consecutive) to complete the learning phase. Once participants met this learning criterion, a transfer phase was initiated. As in Data-set 1, the transfer phase was composed of 5 blocks of 16 stimuli.

Ordering of stimuli. The experiment was characterized by a full factorial design. Three factors were used, each one having two levels: a within-category order manipulation (rule-based vs. similarity-based), a between-category order manipulation (interleaved vs. blocked), and a manipulation of order across blocks (variable vs. constant). The combination of these types of order formed eight conditions (e.g., “rule-based + interleaved + constant”, etc.). The number of participants assigned to each condition is given in Table 1. In the interleaved order, categories were strictly alternated (i.e., *ABABABAB*), while in the blocked order, categories were strictly blocked (i.e., *AAAABBBB* or *BBBBAAAA*). As described above, in the rule-based order stimuli belonging to a same category were presented following a “principal rule plus exceptions” structure, whereas the similarity-based order maximized the similarity between immediately contiguous examples. In the variable manipulation across blocks, the sequence of stimuli varied from one block to another, while in the constant manipulation across blocks, the unique sequence was presented across blocks. For further details we refer the reader to (Mezzadri et al., 2022).

Results

We first present the numerical simulations of accuracy of the parameter estimation and reliability of the method. Then, we present the results of the inference method applied to ALCOVE and Component-cue on both data-sets. Finally, we investigate whether performance of the models is related to the order to which participants were assigned.

Table 1

Number of participants assigned to each of the 8 conditions of Data-set 2.

	Rule-based		Similarity-based	
	Constant	Variable	Constant	Variable
Interleaved	16	14	13	15
Blocked	17	15	21	19

Numerical simulations

Parameter estimation

Numerical simulations were conducted to assess the quality of the parameter estimation, as a function of the size of the data-set. The accuracy of the estimates was highly dependent on whether the exponential or linear version of the models was considered. Parameters of the exponential version were accurately estimated when the size of the data-set was equal to or greater than 80 blocks. On the contrary, parameters of the linear version were overall accurately estimated when the size of the data-set was greater than 160 blocks, with the exception of parameter b (for further details, see Chapter 5 of (Mezzadri, 2020)). Although an accurate estimation of the parameters required data-sets of large size, the predicted probabilities of the models were accurate enough with smaller data-sets (30-40 blocks). Since we were not interested in accurately estimating the parameters of the models but only their classification probabilities, data-sets with 30-40 blocks were judged adequate. Both our data-sets met this condition, thus guarantying an accurate estimation of the predictions of the models. Moreover, since the learning phase of single participants lasted 30 blocks on average, this allowed us to fit participants individually.

Hold-out reliability

Numerical simulations were conducted to assess the reliability of our method. In these simulations, the learning models were used to generate a set of artificial data. These artificial data-sets were then used to determine whether the inference method was able to detect the model with which the data-sets were generated. These steps were iterated 100 times to give a statistical significance to the analysis. The results of the numerical simulations are shown in Figure 4. The graph shows the percentage of times that the simulated data-sets were actually generated by the model that was selected by the method (i.e., the model reaching the lowest evaluation with either the SSD or likelihood criteria). Both criteria gave similar results.

Both the exponential and linear versions of ALCOVE were identified as ALCOVE 81-86% of the time; however, only 56-60% of the time they were identified with the correct version. The exponential version of Component-cue was identified as Component-cue 75-78% of the time, and as Component-Cue^E 64-65% of the time. Finally, the linear version of Component-cue was the most recognizable model with a correct identification of almost 90%. To summarize, our simulations ensure that the model characterized by the lowest evaluation is the model underlying the data (regardless of the version) with a probability of 75-78% at least. Moreover, only the linear version of Component-cue is recognizable with a high probability (almost 90%). The identification of the other versions is not guaranteed with a high probability.

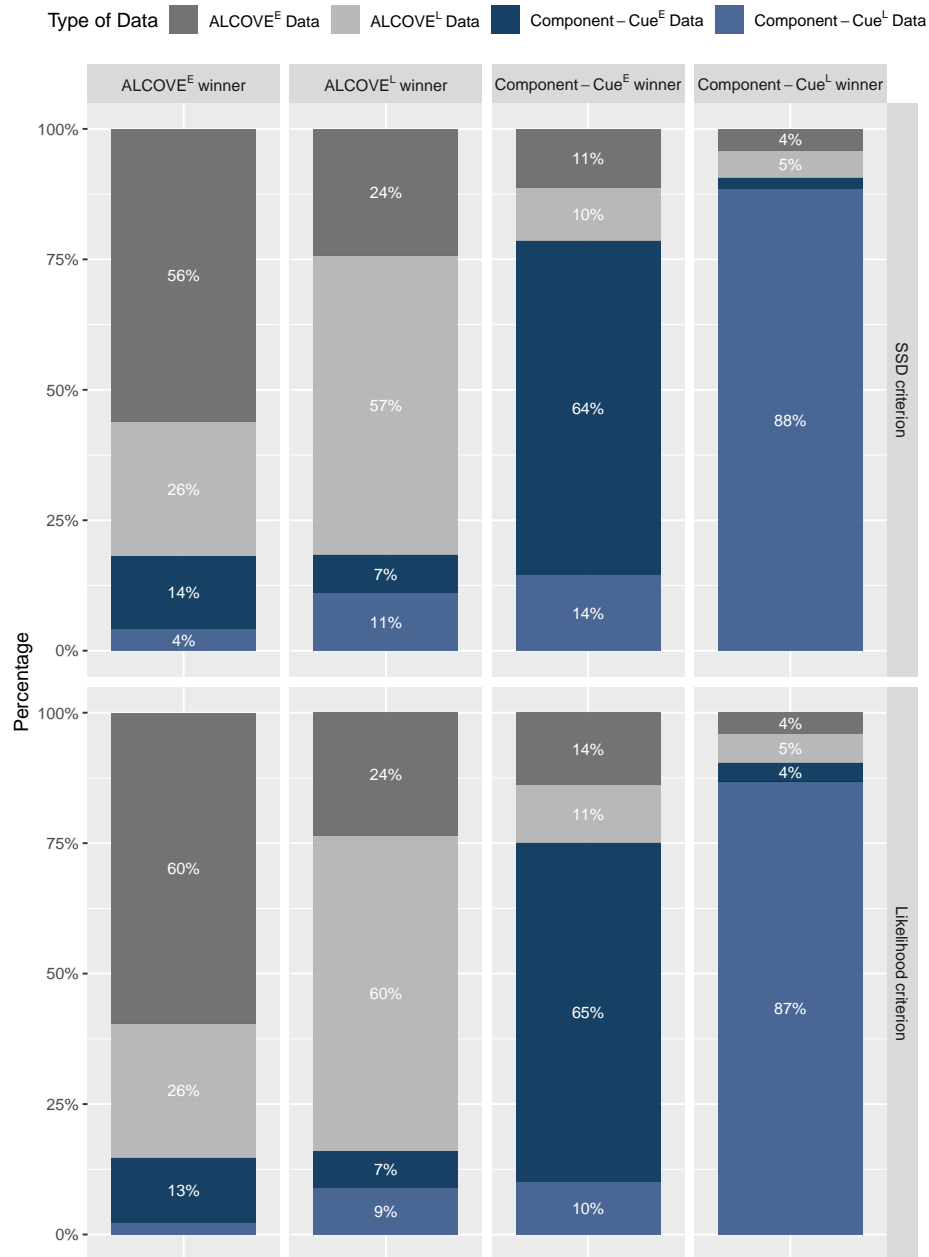


Figure 4

Results of the numerical simulations assessing the reliability of the method. The graph shows the percentage of times that the simulated data-sets were actually generated by the model with the lowest error (using the SSD or likelihood criteria). A total of $20 \text{ iterations} \times 43 \text{ participants} = 860$ hold-out methods were performed. The hold-out method was applied to each participant, separately. Models were fit on the same sequence of stimuli used in Data-set 1: training was performed on the learning phase, while testing was performed on the transfer phase. The gradient descent algorithm in the MLE was performed 10 times.

Table 2

Number of participants that were removed from the analysis, for each condition of Data-set 2. The number of participants who did not reach the learning criterion is shown on the left of “|”, while the number of participants who incorrectly classified more than 25% of the training items during transfer is shown on the right of “|”. For clarifications about the mentioned presentation orders, see Section [Overview of the data-sets](#).

	Rule-based		Similarity-based	
	Constant	Variable	Constant	Variable
Interleaved	2 3	0 4	0 1	1 6
Blocked	0 4	1 5	0 9	2 4

Model-fitting results

Here, we present the results of the application of the hold-out cross-validation method to Data-set 1 and Data-set 2. Since we were interested in studying how participants learned and remembered the categories, those who did not meet this criterion were removed from the analysis. In Data-set 1, one participant (in the similarity-based order) did not meet the learning criterion, whereas in Data-set 2, 6 participants did not meet the learning criterion. Also, 36 participants incorrectly classified more than 25% of the training items during transfer and were then removed from the analysis (for details on which condition they were assigned to see Table 2). Regarding trials in which participants did not classify stimuli on time (amounting to 1.4% in Data-set 2; participants in Data-set 1 always classified stimuli on time), one of the two categories was randomly selected to facilitate modeling.

Data-set 1

Figure 5 (on the top) shows the results of the application of the hold-out method to Data-set 1, with the transfer phase as the testing set. Each participant was fit separately. The graph shows the number and percentage of participants who were best fit by the various learning models, depending on the evaluation criteria. Component-cue best performed on 63-66% of the participants, with a dominance of the linear version with the SSD criterion and a dominance of the exponential version with the likelihood criterion. Simulations ensured us that the model underlying the responses of the participants who were best fit by Component-cue was actually Component-cue with a probability of 75-78%. Moreover, when the best-fitting model was ALCOVE or Component-cue^L the probability increased to 81-86% or 89-93%, respectively.

Data-set 2

Figure 5 (in the middle and on the bottom) shows the results of the application of the hold-out method to Data-set 2, with either the transfer phase or the unsupervised blocks of the learning phase as the testing set. Each participant was fit separately. The graph shows the number and percentage of participants who were best fit by the learning models, depending on the evaluation criteria. In the graph in the middle, models were trained on the supervised blocks of the learning phase and tested on the transfer phase. Approximately half of the participants were best fit by

ALCOVE (44-48%) and half of the participants were best fit by Component-cue (52-56%). In the graph on the bottom, models were trained on the supervised blocks of the learning phase and tested on the unsupervised blocks of the same phase. This time, the majority of the participants (75-81%) was best fit by Component-cue, with a dominance of the exponential version. Again, simulations ensured us that these results are liable with a probability of 75-78%.

Relation between models and within-category orders

Here, we investigate a potential connection between the two strategies at play in the models (a complex rule-based strategy in Component-cue and a similarity-based strategy in ALCOVE) and the within-category orders used to present stimuli (rule-based and similarity-based orders). Again, one plausible hypothesis is that a model integrating a mechanism X should be favored by an order inspired by X . Another way to put it is that it would be paradoxical for a model to show a preference for a type of presentation which a priori would not favor its implemented mechanisms. To investigate this hypothesis, we analyzed *i*) whether the number of participants that were best fit by a specific model was related to the within-category order assigned to these participants (in “Distribution of participants analysis”), and *ii*) whether the generalization patterns of the model that best fit our data were related to the within-category order in which stimuli were presented (in “Generalization patterns analysis”). The description of the analysis on generalization patterns will be preceded by a test of the difference of performance between participants in the rule-based and similarity-based orders. This additional test will serve as a baseline for our analysis on generalization patterns.

Distribution of participants analysis

Data-set 1. Table 3 (on the top) shows the number of participants in Data-set 1 whose responses were best predicted by either Component-cue or ALCOVE, as a function of the within-category order and evaluation criterion. With both criteria, Component-cue best fitted a higher number of participants in the rule-based order as compared to the similarity-based order. Conversely, ALCOVE best fitted a higher number of participants in the similarity-based order as compared to the rule-based order. Moreover, participants assigned to the rule-based order were overall best fit by Component-cue. A Fisher’s exact test of independence was separately performed on the SSD and likelihood tables to assess whether the relation that emerged between models and orders was significant. None of the two tests were found significant (p-value = 0.06 with the SSD and p-value = 0.53 with the likelihood). The striking difference between the two p-values might be due to the small sample of the data-set.

Data-set 2. Table 3 (on the bottom) shows the number of participants in Data-set 2 whose responses were best predicted by either Component-cue or ALCOVE, as a function of the within-category order and evaluation criterion. We considered the results of the analysis in which models were trained on the supervised blocks of the learning phase and tested on the transfer phase. Again, Component-cue best fitted more participants in the rule-based order than in the similarity-based order, while ALCOVE best fitted more participants in the similarity-based order than in the rule-based order. Moreover, participants assigned to the rule-based order were overall best fit by Component-cue, whereas participants assigned to the similarity-based order were overall best fit by ALCOVE. Although the relation between models and orders was more visible than before, the Fisher’s exact test of independence was not significant (p-value = 0.09 with the SSD and p-value = 0.14 with the likelihood).

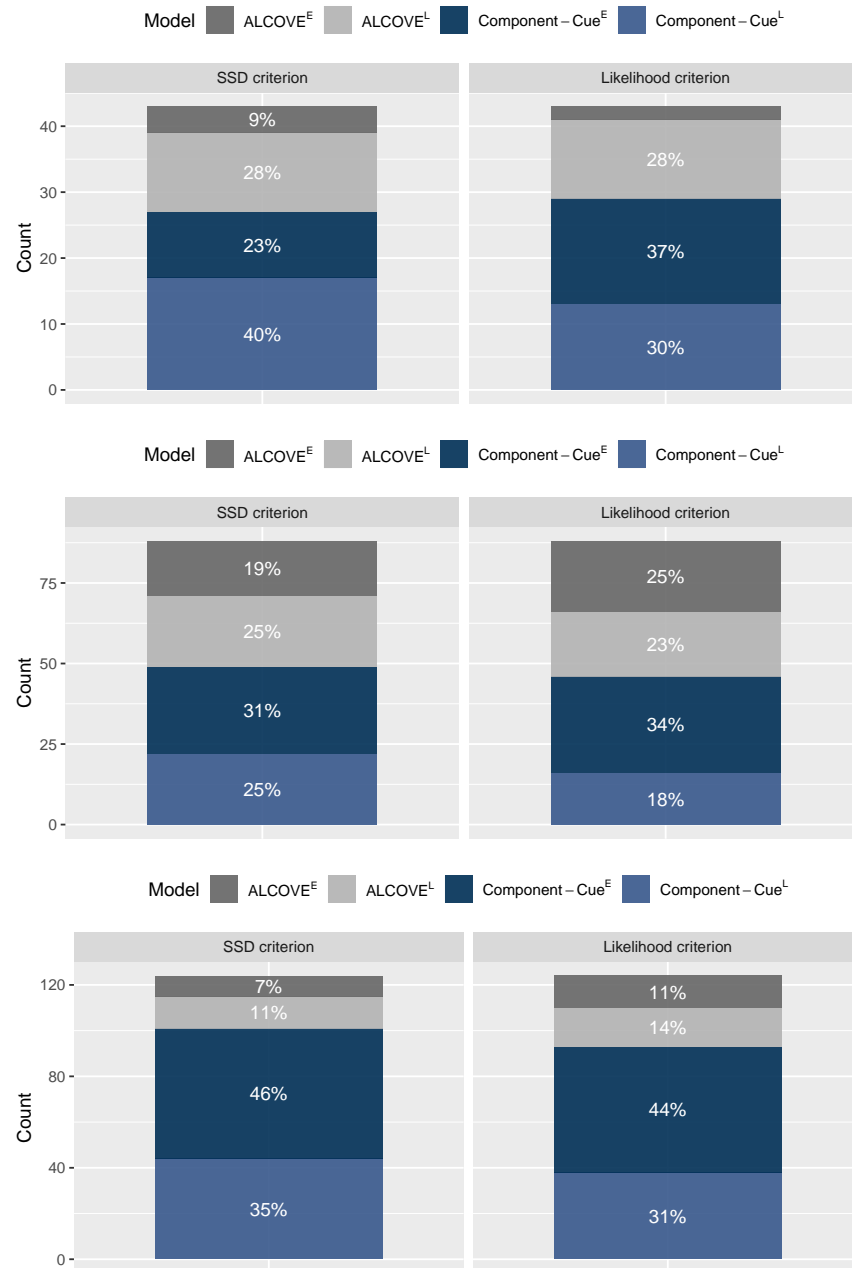


Figure 5

Application of the hold-out method on Data-set 1 (on the top) and Data-set 2 (in the middle and on the bottom). The graphs show the number and percentage of participants that were best fit by the learning models, as a function of the models and evaluation criteria. On the top, models were trained on the learning phase and tested on the transfer phase (Data-set 1). In the middle, models were trained on the supervised blocks of the learning phase and tested on the transfer phase (Data-set 2). On the bottom, models were trained on the supervised blocks of the learning phase and tested on the unsupervised blocks of the learning phase (Data-set 2).

Table 3

Number of participants in Data-set 1 (on the top) and Data-set 2 (on the bottom) whose responses were best predicted by either Component-cue or ALCOVE, as a function of the within-category order (rule-based vs. similarity-based) and evaluation criterion (SSD vs. Likelihood).

Type of order	SSD		Likelihood	
	ALCOVE	Component-cue	ALCOVE	Component-cue
<i>Data-set 1</i>				
Rule-based	5	17	6	16
Sim.-based	11	10	8	13
<i>Data-set 2</i>				
Rule-based	15	28	17	26
Sim.-based	24	21	25	20

Additional test on generalization patterns

We previously mentioned that Data-set 1 was used in (Mathy & Feldman, 2016) to show the influence of within-category presentation order on generalization patterns. Here, we present an additional test confirming this result. This test represents a baseline for our next analysis. We considered participants’ generalization patterns by computing the proportion of time (across transfer) that each participant classified the transfer items into category A . By definition, participants adopting a rule-based strategy would classify new stimuli on the basis of the main rule (which for Data-set 1 is all gray items belong to category A and all blue items belong to category B , and for Data-set 2 is all striped items belong to category A and all plain items belong to category B). The use of this strategy would produce the following putative generalization pattern (see Figure 3):

$$(\mathbb{P}(A | T_1), \dots, \mathbb{P}(A | T_7)) = (1, 1, 0, 0, 0, 0, 1),$$

where $\mathbb{P}(A | T_i)$ is the probability to classify transfer item T_i into category A (for $i = 1, \dots, 7$; we had 7 transfer items). Inversely, participants adopting a similarity-based strategy would classify new stimuli on the basis of their similarity to stored items. The use of this strategy would produce the following putative generalization pattern (see Figure 3; only the category assignment of the closest items was used):

$$(\mathbb{P}(A | T_1), \dots, \mathbb{P}(A | T_7)) = \left(\frac{1}{4}, \frac{1}{2}, 0, \frac{1}{2}, \frac{1}{4}, \frac{1}{2}, 1 \right).$$

Comparing these two stereotypical generalization patterns, participants adopting a rule-based strategy would classify items T_1 and T_2 , items T_3 and T_7 , and items T_4 , T_5 and T_6 into category A with, respectively, a higher, an equal, and a lower probability as compared to participants adopting a similarity-based strategy. This means that, if participants’ generalization patterns are projected on the vector $v = (1, 1, 0, -1, -1, -1, 0)$, participants following a rule-based strategy have higher projection values than participants following a similarity-based strategy.

Therefore, the influence of presentation order on generalization patterns can be studied in a straightforward manner by analyzing the projections of participants’ generalization patterns

on vector v , as a function of the presentation order. A difference in location between participants in the rule-based and similarity-based orders would mean a difference in generalization patterns. This difference in location was assessed running a one-sided Wilcoxon-Mann-Whitney test (we tested whether participants following a rule-based order had higher projection values than those following a similarity-based order). The use of the Wilcoxon-Mann-Whitney test allowed us to avoid making hypotheses about the distribution underlying the data. The test turned out significant for both data-sets (p-value = 0.021 for Data-set 1 and p-value = 0.022 for Data-set 2), showing that the generalization patterns of participants in the rule-based order were closer to a rule-based classification pattern as compared to participants in the similarity-based order.

Generalization patterns analysis

This analysis aims at investigating whether the order assigned to participants (rule-based vs. similarity-based) was related to the model that best fit their responses (ALCOVE vs. Component-cue). Similarly to the previous analysis, the generalization patterns provided by the best-fitting model were projected on the vector $v = (1, 1, 0, -1, -1, -1, 0)$. Two tests were conducted: a first test assessing the difference in location between the projections of ALCOVE and those of Component-cue, and a second test assessing the difference in location between the projections of participants in the rule-based order and those of participants in the similarity-based order. Both made use of a one-sided Wilcoxon-Mann-Whitney test. Because we expected to find that the closest projections to a rule-based generalization pattern would have been those of Component-cue and those of the participants in the rule-based order, a one-sided test was preferred to a two-sided test.

Data-set 1. The first test was significant (p-value = 0.0002), showing that the projections of the generalization patterns of Component-cue were higher than the projections of the generalization patterns of ALCOVE. This means that, when considering the best-fitting model, the generalization patterns of Component-cue were more consistent with a rule-based retrieval than those of ALCOVE. The second test fell short of significance (p-value = 0.059), showing that the projections of the generalization patterns of participants in the rule-based order were slightly higher than the projections of the generalization patterns of participants in the similarity-based order. This means that the best-fitting models were not clearly able to reproduce the difference in generalization patterns found in the data between participants in the rule-based and similarity-based orders.

Data-set 2. As Data-set 1, only the first test was significant (p-value = 0.0005 for the first and p-value = 0.373 for the second). Again, the generalization patterns of Component-cue were more consistent with a rule-based retrieval than those of ALCOVE, when considering the best-fitting model. However, the best-fitting models were not able to reproduce the fact that different types of presentation order created a distortion in the representation of the categories.

Discussion

In the last three decades, research in categorization has seen a rapid evolution of models of category learning and representation (Carvalho & Goldstone, 2022; Lee & Webb, 2005; Love et al., 2004; Mezzadri et al., 2022; Nosofsky & Palmeri, 1998). However, little effort has been directed toward the promotion of a rigorous method for comparing learning models (Pitt et al., 2002), and a common testing ground is still lacking. Our study attempts to address this question by presenting a general inference method for the selection of learning models.

Our main contribution includes the promotion of a general method for fitting learning models to data. Here, we propose the use of a cross-validation method (hold-out) as a better technique than classical statistical criteria to account for dependent data and small samples. Training and testing sets in the hold-out method were appropriately selected to suit our objective. To study performance during transfer, models were trained during learning and tested during transfer, whereas to study learning progressions, models were trained on supervised learning blocks and tested on unsupervised learning blocks. Because the learning phase is generally composed of supervised blocks exclusively, an experiment that alternates blocks of supervised learning with blocks of unsupervised learning was specifically designed to fit our purposes. Also, numerical simulations assessing the accuracy of the parameter estimation allowed us to apply the method to individual data.

Here, we summarize and generalize a series of good practices, that we hope will serve as guidelines for future studies. A first good practice is to make use of cross-validation techniques to test learning models. Training and testing models on different subsets allows the respect of the structure of dependency within the data, while avoiding the risk of over-fitting them. Our results suggest the use of the hold-out method as an adequate trade-off between reliability, complexity of its application, and computational cost. A second good practice is to evaluate how well parameters (or alternatively, classification probabilities) are estimated, as a function of the size of the dataset. This analysis allows researchers to determine whether models can be fit to individual or collective data. A third good practice is to select the same criterion for estimating the parameters and evaluating the models concurrently. Finally, a fourth good practice is to study whether the selected cross-validation method is able to identify the model underlying the artificial data that were generated with it.

Another contribution is the application of our inference method to compare two common category learning models: ALCOVE and Component-cue. These models were not chosen because we consider them as being the most representative of the domain, but because of their similar underlying structure. These models implement alternative strategies, while sharing a similar neuron network structure. ALCOVE implements a similarity-based strategy, whereas Component-cue rather implements a (possibly complex) rule-based strategy. By fitting these models to two datasets, we found that during transfer (in both Data-set 1 and Data-set 2) almost half of the participants were best fit by ALCOVE, while during learning (in Data-set 2; this analysis on Data-set 1 was not possible) the majority of participants were best fit by Component-cue. Our numerical simulations ensured the reliability of these results with a high confidence (75-78% at least). A complex rule-based strategy was preferred by participants during learning, while both a complex rule-based and a similarity-based strategies were approximately equally used during transfer. However, the nature of the task might have favored the use of a complex rule-based strategy during learning.

An additional explanation for the difference in the fitting performance of the studied models can be found in neurocomputational theories. It is widely accepted that learning is mediated by at least two memory systems: an explicit system that depends largely on the prefrontal cortex and uses rule-governed mechanisms, and an implicit procedural system that depends on the basal ganglia (Poldrack & Packard, 2003; Ashby & Valentin, 2017). Paul & Ashby (2013) have hypothesized a one-way interaction between these two systems, in which the response generated by the explicit system is communicated back to the procedural system. More recently, the procedural system has been shown to be equivalent to a neural integration of similarity-based exemplar theory (Ashby & Rosedahl, 2017), which is what ALCOVE was designed to be. Therefore, the fact

that Component-Cue performed well during learning, whereas ALCOVE performed best during transfer may be related to these ideas.

A last contribution includes the investigation of a putative relation between best-fitting models (ALCOVE and Component-cue) and within-category orders (rule-based and similarity-based). This investigation was driven by the hypothesis that an environment fitting the internal mechanism of a model should facilitate the extraction of the categories. Therefore, a rule-based order should be beneficial for models implementing a rule-based or a complex rule-based strategy (such as Component-cue), whereas a similarity-based order should be beneficial for models implementing a similarity-based strategy (such as ALCOVE). A first factor we investigated was the number of participants in the rule-based and similarity-based orders that were best fit by the chosen category learning models. Although participants in the rule-based order were best fit by Component-cue (in both data-sets) and participants in the similarity-based order were best fit by ALCOVE (in Data-set 2), the difference was not significant (in both data-sets). A second factor we investigated is the generalization patterns provided by the best-fitting model, as a function of presentation orders. The results showed that the generalization patterns of Component-cue were closer to a rule-based classification than the generalization patterns of ALCOVE. An additional analysis showed that the models that best fit participants' performance were not able to reproduce the difference in generalization patterns assessed during transfer between participants in the rule-based and similarity-based orders. To summarize, we found some evidence pointing toward a relation between models and presentation order; however, further investigations are necessary to evaluate the consistency of this result.

Limitations and Perspectives

Our study only compared two category learning models, without including other relevant models that implement other learning strategies (Ashby et al., 1998; Erickson & Kruschke, 1998; Gluck & Bower, 1988; Kruschke, 1992; Kruschke & Johansen, 1999; Love et al., 2004; Nosofsky & Palmeri, 1998). Also, the categories we studied (the 5-4 category set) are characterized by a clear rule-plus-exceptions structure. The use of such a structure might have promoted the adoption of a rule-based strategy over a similarity-based one. To overcome these limitations, the present study should be extended to a larger variety of models and categories.

In machine learning, the hold-out method is used to identify discrepancies between feedback and predictions of models (Kopper et al., 2020; Yadav & Shukla, 2016). In our article, the method was used in a different spirit, with the idea that models are trained using feedback whereas they can be tested using participants' responses when no feedback is provided. Therefore in machine learning, feedback serves both as a training tool and a testing tool, while in our context two different tools were used with the purpose of training and testing the models (feedback as training tool and participants' responses when no feedback is provided as testing tool). Although this adopted strategy seems to give good results on simulated data (see Figure 4 page 12), a more rigorous mathematical formalization of this new inference method is needed. Also, there are experiments where participants are never deprived of feedback. This is especially true in cognitive tasks for non-human animals in which reward for a correct behavior is always given. In a current work under redaction (Moongathottathil-James et al., 2021), we are thus using a different approach for applying the hold-out method to learning models on data with constant feedback. In this case, we used the last part of the learning phase as the testing set and we allowed parameters to be updated during testing as well.

Finally, a word of caution has to be said about the combination of both feedback and no-feedback trials. Previous studies have found that immediate vs. delayed feedback affects the accuracy of responding and the distribution of best fitting models in information-integration category-learning tasks (Maddox et al., 2003; Maddox & David, 2005). Moreover, it has been shown that disrupting feedback processing time (short vs. long) interferes with rule-based category learning (Maddox et al., 2004). Although evidence comparing our design with a design characterized by uninterrupted feedback is needed, it is worth mentioning that the inclusion of no-feedback trials in the learning phase of our task might have influenced the strategy used by participants.

Acknowledgements

The present work was supported by the French government, through the UCAJedi and 3IA Côte d’Azur Investissements d’Avenir managed by the National Research Agency (ANR-15-IDEX-01 and ANR-19-P3IA-0002), directed by the National Research Agency with the ANR project ChaMaNe (ANR-19-CE40-0024-02) and by the interdisciplinary Institute for Modeling in Neuroscience and Cognition (NeuroMod) of the Université Côte d’Azur.

Author contributions

The inference method was developed by G.M. with contributions from P.R.-B., T.L. and F.M. Study design and data analysis were performed by G.M. The article was drafted by G.M. and critical revisions were provided by P.R.-B., T.L. and F.M. All authors approved the final version of the manuscript for submission.

Conflict of interest

The authors declare no conflict of interest with respect to their authorship or the publication of this article.

Open Practices Statement

The data-sets and computer code used in the current study (including the code for reproducing tables and figures) are publicly available in Open Science Framework at https://osf.io/5jn24/?view_only=b7fd79c283e54e098a27ecabe9e1346f.

References

- Akaike, H. (1998). *Information Theory and an Extension of the Maximum Likelihood Principle*. Springer New York.
- Aldrich, J. (1997). R.A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3), 162–176.
- Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16, 125–127.
- Arifovic, J., & Ledyard, J. (2011). A behavioral model for mechanism design: Individual evolutionary learning. *Journal of Economic Behavior & Organization*, 78, 374–395.

- Ashby, F., Alfonso-Reese, L., Turken, A., & Waldron, E. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481.
- Ashby, F. G., & Rosedahl, L. (2017). A neural interpretation of exemplar theory. *Psychological Review*, *124*(4), 472.
- Ashby, F. G., & Valentin, V. V. (2017). Chapter 7 - multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen, & C. Lefebvre (Eds.) *Handbook of Categorization in Cognitive Science (Second Edition)*, (pp. 157–188). San Diego: Elsevier, second edition ed.
- Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 185–207.
- Bower, G., Clark, M., Lesgold, A., & Winzenz, D. (1969). Hierarchical retrieval schemes in recall of categorized word lists. *Journal of Verbal Learning and Verbal Behavior*, *8*, 323–343.
- Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*, 481–495.
- Carvalho, P. F., & Goldstone, R. L. (2022). A computational model of context-dependent encodings during category learning. *Cognitive Science*, *46*(4), e13128.
- Cawley, G., & Talbot, N. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.
- Claeskens, G., & Hjort, N. (2007). *Model Selection And Model Averaging*. Cambridge.
- Dreyfus, S. E. (1990). Artificial neural networks, back propagation, and the kelley-bryson gradient procedure. *Journal of Guidance Control and Dynamics*, *13*, 926–928.
- Elio, R., & Anderson, J. (1981). The effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 397–417.
- Elio, R., & Anderson, J. R. (1984). The effects of information order and learning mode on schema abstraction. *Memory & Cognition*, *12*, 20–30.
- Erickson, M., & Kruschke, J. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140.
- Gluck, M., & Bower, G. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, *27*, 166–195.
- Högden, F., Stahl, C., & Unkelbach, C. (2019). Similarity-based and rule-based generalization in the acquisition of attitudes via evaluative conditioning. *Cognition and Emotion*, *34*, 105–127.
- Konishi, S., & Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer.
- Kopper, A., Karkare, R., Paffenroth, R., & Apelian, D. (2020). Model selection and evaluation for machine learning: Deep learning in materials processing. *Integrating Materials and Manufacturing Innovation*, *9*, 287 – 300.

- Kruschke, J., & Johansen, M. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1083–1119.
- Kruschke, J. K. (1992). Alcové: an exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22–44.
- Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, 12, 605–621.
- Lemaire, B., & Portrat, S. (2018). A computational model of working memory integrating time-based decay and interference. *Frontiers in Psychology*, 9, 416.
- Love, B., Medin, D., & Gureckis, T. (2004). Sustain: a network model of category learning. *Psychological Review*, 111, 309–332.
- Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(4), 650.
- Maddox, W. T., Ashby, F. G., David, A., & Pickering, A. D. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition*, 32(4), 582–591.
- Maddox, W. T., & David, A. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(1), 100.
- Malem-Shinitzki, N., Opper, M., Reich, S., Schwetlick, L., Seelig, S. A., & Engbert, R. (2020). A mathematical model of local and global attention in natural scene viewing. *PLOS Computational Biology*, 16(12), 1–21.
- Mathy, F., & Feldman, J. (2009). A rule-based presentation order facilitates category learning. *Psychonomic Bulletin & Review*, 16, 1050–1057.
- Mathy, F., & Feldman, J. (2016). The influence of presentation order on category transfer. *Experimental Psychology*, 63, 59–69.
- Medin, D., & Bettger, J. (1994). Presentation order and recognition of categorically related examples. *Psychonomic Bulletin & Review*, 1, 250–254.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85, 207–238.
- Mezzadri, G. (2020). Statistical inference for categorization models and presentation order. PhD thesis on webpage at tel.archives-ouvertes.fr/tel-03219311.
- Mezzadri, G., Reynaud-Bouret, P., Laloë, T., & Mathy, F. (2022). An order-dependent transfer model in categorization. *Journal of Mathematical Psychology*, 107, 102634.

- Moongathottathil-James, A., Mezzadri, G., Sargolini, F., Reynaud-Bouret, P., Bethus, I., & Muzy, A. (2021). Do rats learn by paths or by turns when exploring a maze? determining and predicting automatically the actions involved in the learning process. Preprint on webpage at hal.archives-ouvertes.fr/hal-03093527.
- Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.
- Myung, I. J., & Pitt, M. A. (1998). Issues in selecting mathematical models of cognition. In J. Grainger & A. M. Jacobs (Eds.), *Scientific Psychology Series. Localist Connectionist Approaches To Human Cognition*, (p. 327–355).
- Nosofsky, R., Gluck, M., Palmeri, T., Mckinley, S., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of shepard, hovland, and jenkins (1961). *Memory & Cognition*, *22*, 352–369.
- Nosofsky, R., Kruschke, J., & McKinley, S. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 211–233.
- Nosofsky, R. M., & Palmeri, T. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychon. Bull. Rev.*, *5*, 345–369.
- Nosofsky, R. M., Sanders, C., Zhu, X., & Mcdaniel, M. (2018). Model-guided search for optimal natural-science-category training exemplars: A work in progress. *Psychonomic Bulletin & Review*, *26*, 48–76.
- Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2017). Tests of an exemplar-memory model of classification learning in a high-dimensional natural-science category domain. *Journal of Experimental Psychology: General*, *147*, 328–353.
- Novikov, D., Korepanov, V., & Chkhartishvili, A. (2018). Reflexion in mathematical models of decision-making. *International Journal of Parallel, Emergent and Distributed Systems*, *33*, 1–17.
- Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language*, *55*(4), 601–626.
- Palmeri, T. (1999). Learning categories at different hierarchical levels: A comparison of category learning models. *Psychonomic Bulletin & Review*, *6*, 495–503.
- Paul, E. J., & Ashby, F. G. (2013). A neurocomputational theory of how explicit learning bootstraps early procedural learning. *Frontiers in Computational Neuroscience*, *7*, 177.
- Pitt, M., Myung, I., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.
- Poldrack, R. A., & Packard, M. G. (2003). Competition among multiple memory systems: converging evidence from animal and human brain studies. *Neuropsychologia*, *41*(3), 245–251.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408.

- Roth, A. E., & Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, 8, 164–212.
- Sanders, C., & Nosofsky, R. (2020). Training deep networks to construct a psychological feature space for a natural-object category domain. *Computational Brain & Behavior*, (pp. 1–23).
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Stone, M. (1974). Cross-validated choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36, 111–133.
- Yadav, S., & Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, (pp. 78–83).