Behavior Research Methods (2024) 56:2549–2568
https://doi.org/10.3758/s13428-023-02166-6

How to fit transfer models to learning data: a segmentation/clustering approach

Giulia Mezzadri[1] · Thomas Laloë[2] · Fabien Mathy[3] · Patricia Reynaud-Bouret[2]

Accepted: 6 June 2023 / Published online: 17 July 2023

Abstract
Although transfer models are limited in their ability to evolve over time and account for a wide range of processes, they have repeatedly shown to be useful for testing categorization theories and predicting participants' generalization performance. In this study, we propose a statistical framework that allows transfer models to be applied to category learning data. Our framework uses a segmentation/clustering technique specifically tailored to suit category learning data. We applied this technique to a well-known transfer model, the Generalized Context Model, in three novel experiments that manipulated ordinal effects in category learning. The difference in performance across the three contexts, as well as the benefit of the rule-based order observed in two out of three experiments, were mostly detected by the segmentation/clustering method. Furthermore, the analysis of the segmentation/clustering outputs using the backward learning curve revealed that participants' performance suddenly improved, suggesting the detection of an "eureka" moment. Our adjusted segmentation/clustering framework allows transfer models to fit learning data while capturing relevant patterns.

Keywords Categorization · Segmentation/Clustering · Category transfer models · Generalized Context Model (GCM) · Rule-based versus similarity-based presentation order

Introduction

Cognitive sciences have seen significant progress due to the conception and use of computational models (Polk and Seifert, 2002; Sun, 2008; Busemeyer and Diederich, 2010; Lieto, 2021). This is particularly true in categorization (Pothos and Wills, 2011; Wills, 2013), where models have been developed to better understand underlying mechanisms (Reed, 1972; Hintzman, 1984; Nosofsky et al., 1994; Love et al., 2004; Kruschke, 2008) and, more recently, order effects (Carvalho and Goldstone, 2022; Mezzadri et al., 2022c). Following the learning vs. generalization distinction, computational models can be grouped into learning and transfer models. Learning refers to the formation of the categories through a trial-and-error process, whereas transfer refers to the ability to classify new stimuli. Learning models have the ability to adapt their predictions over time, which allows them to perform equally well on learning and transfer data. Examples of learning models are the Configural-Cue network model of classification learning (Gluck and Bower, 1988), the Attention Learning COVEring map (ALCOVE) model of categorization (Kruschke, 1992), the Rule-plus-exception (RULEX) model of classification learning (Nosofsky et al., 1994), and SUSTAIN (Love et al., 2004). By contrast, transfer models tend to produce predictions that do not evolve significantly over time. This limitation often restricts their usage to transfer only. Examples of transfer models are the Generalized Context Model (Nosofsky, 1986) and the Ordinal General Context Model (Mezzadri et al., 2022c). We refer to Supplementary material A for a visual comparison of the performance between learning and transfer models as a function of time.

Although transfer models account for fewer cognitive processes than learning models (focusing on generalization rather than both learning and generalization processes), they have been shown to be useful in accurately predicting participants' performance across a variety of contexts

✉ Giulia Mezzadri
gm3026@columbia.edu

[1] Cognition and Decision Lab, Columbia University, New York, US

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

[3] Laboratoire Bases, Corpus, Langage UMR CNRS 7320, Université Côte d'Azur, Nice, France

(Nosofsky et al., 2018, 2017; Rehder and Hoffman, 2005; Rouder and Ratcliff, 2004; Sanders and Nosofsky, 2020; Smith and Minda, 2000). They also have been used to implement theories of cognitive processes, such as models based on exemplars vs. prototypes (Minda and Smith, 2002). For instance, Nosofsky et al. (2018) have recently tested the ability of a well-known exemplar model of categorization, the Generalized Context Model (GCM), to predict classification of rocks. In this study, the authors showed that GCM was capable of providing precise quantitative predictions for various training conditions applied to diverse rock categories.

We here propose a statistical framework that allows learning data to be fit by transfer models with the aim of extending the application of transfer models to category learning. The advantages of applying transfer models to learning data are numerous. First, this framework can allow transfer models to be applied to experiments that lack a transfer phase. This is the case for certain classification studies in which transfer items are not conceived (Ashby and Maddox, 1992; Feldman, 2000, 2003), but also the case in cognitive tasks designed for non-human animals in which reward for a correct behavior is always provided (James et al., 2022). It is particularly useful when the aim is to identify learning stages, rather than using a predetermined model of the learning phase.

Second, this framework can allow a fruitful use of the learning phase of a classification task. As mentioned above, transfer models are best suitable for fitting transfer data. Yet, the transfer phase of a categorization experiment is generally short (amounting to a few blocks). Therefore, considering that a portion of the transfer phase is used for estimating the parameters, there are generally a few blocks left for testing the predictions of the model. Using the framework of the present study, parameter estimation can be performed on learning data, letting one to assess models on the whole data set.

Finally, this method can give researchers the choice of fitting participants' generalization patterns individually or collectively. The main obstacle to individually fit participants in the transfer phase lies in the greater amount of individual data needed to accurately estimate the parameters (Mezzadri et al., 2022a). Since the use of our method would allow the estimation of the parameters on the learning phase (which generally includes enough observations to accurately estimate the parameters), a participant-by-participant fit would then be possible.

The statistical framework that we propose is based on a segmentation/clustering approach (Picard et al., 2007), originally applied to DNA data (Davies et al., 2005). The segmentation/clustering model combines a segmentation model with a mixture model. The former divides the data into a finite number of segments and the latter assigns a label to each segment. In the case of classification data, each label is associated with a specific learning behavior/phase (e.g., random classification, perfect classification, etc.). From now on, the term "behavior" is preferred to "label" to facilitate interpretation in categorization. Partitioning the data into segments allows transfer models to adapt their predictions to the learning path of each participant, making it an individual fit. On the other hand, assigning a behavior to each segment allows the comparison between participants. Indeed, the method benefits from all of the observations in estimating the parameters of each behavior making the estimation robust and interpretable through all individuals. To our knowledge, such a method has never been applied to cognitive models.

The use of a segmentation/clustering framework was preferred to simpler segmentation methods for two reasons. First, segmentation methods do not allow the attribution of a behavior to segments, which makes comparisons among participants' learning paths more difficult. While segmentation methods only rely on an individual fit, the segmentation/clustering framework is characterized by a dual individual/collective fit. As mentioned above, the method recovers the sequence of behaviors through an individual fit, but identifies these behaviors based on the set of collective data points (allowing comparisons among subjects). Second, the segmentation/clustering method allows a more accurate estimation of the parameters compared to segmentation methods. Indeed, the number of observations per segment in individuals who rapidly completed the experiment might be too small to accurately estimate the parameters with segmentation methods (Mezzadri et al., 2022c). By contrast, the segmentation/clustering method allows parameters to be estimated on the set of segments associated with a common behavior, which generally includes a higher number of observations.

We here apply the segmentation/clustering method to the Generalized Context Model (Medin and Schaffer, 1978; Nosofsky, 1986). This model can account for a variety of category-learning phenomena, and has served as a general framework for a large number of significant models in categorization (e.g., Anderson, 1991; Kruschke, 1992; Love et al., 2004). Three novel experiments were conducted to evaluate the application of the segmentation/clustering method to the Generalized Context Model. These experiments only involve a learning phase. We manipulated the order of stimuli within a category to obtain variations of performance in the data. We focused on two specific within-category orders: rule-based, in which members of a same category are presented following a "principal rule plus exceptions" structure, and similarity-based, in which members of a same category are arranged in order to maximize the similarity between contiguous stimuli. Research has shown that the rule-based order facilitates learning as compared to the similarity-based order when the category structure itself favors the abstraction of a rule (Elio and Anderson, 1981, 1984; Mathy and Feldman, 2009, 2016; Mezzadri et al., 2022b). Here, these two types of presentation order are studied in various contexts.

In "Segmentation/clustering framework applied to the Generalized Context Model (GCM)", we provide a brief overview of the Generalized Context Model and describe the segmentation/clustering framework. Within this section, we also present the results of the numerical simulations used to validate the performance of the segmentation/clustering method, as well as to optimize the selection of parameters for the number of behaviors and change-points. In "Experiments", we provide descriptions of three novel experiments, while in "Data Analysis" analyze the data collected from these experiments. Finally, in "Results", we present the results obtained from applying the segmentation/clustering technique to the Generalized Context Model across all three data sets.

Segmentation/clustering framework applied to the Generalized Context Model (GCM)

The segmentation/clustering technique (Picard et al., 2007) combines a segmentation model whose purpose is to detect abrupt changes within the data (Hupé et al., 2004; Olshen et al., 2004), with a mixture model which assigns a behavior to each early detected segment. The segmentation model provides a partition of the data into segments while accounting for the ordered structure of the data, whereas the mixture model allows the association of a common behavior to segments with similar features. We first provide a brief description of the Generalized Context Model (GCM), to which the segmentation/clustering technique is applied. Then, we describe the segmentation/clustering model for a fixed number of change-points and behaviors. Finally, we address the selection of the number of change-points and behaviors, and we present numerical simulations assessing the reliability of the segmentation/clustering model. Note that the number of change-points and behaviors remains fixed within a single experiment, but may vary across different experiments.

Overview of the Generalized Context Model (GCM)

To improve readability, all main symbols used in this sub-section and the following sub-sections are listed and explained in Table 1. As our experiments only consider two categories, GCM is formalized specifically for this case. According to GCM (Medin and Schaffer, 1978; Nosofsky, 1986), the probability of classifying a stimulus x as belonging to the set of positive stimuli (i.e., Category $+$) is given by the summed similarities of that stimulus to all positive learning stimuli, divided by the summed similarities of stimulus x to all learning stimuli of both categories (i.e., Category $+$ and Category $-$):

Table 1 Main symbols used in the study as well as their definition

Symbol	Definition
$+/-$	Set of positive/negative stimuli
$\mathcal{S}(\cdot,\cdot)$	Similarity between two stimuli
c	Sensitivity parameter in GCM
$\mathbb{P}_c\left(+\,\middle|\,\cdot\right)$	Probability of classifying an item into Category $+$ as a function of the sensitivity parameter c
$d(\cdot,\cdot)$	Distance between two stimuli
$\mathcal{D}$	Number of dimensions of the psychological space in which stimuli are embedded
ω_i	Attention-weight for dimension i
$\mathcal{L}(\cdot\,;\,c)$	Likelihood function given a sequence of stimuli as a function of the sensitivity parameter c
m	A given participant
$z_1^m,\ldots,z_{n_m}^m$	Sequence of responses given by participant m involving n_m data points
P	Number of behaviors
K	Number of change-points
$\theta=\{\theta_1,\ldots,\theta_P\}$	Values of the sensitivity parameter associated with each behavior
$\tau^m=\{\tau_1^m,\ldots,\tau_K^m\}$	Coordinates of the change-points for participant m
S_k^m	The k-th segment for participant m obtained from the segmentation of the data
α	Parameter for selecting the number of behaviors
β	Parameter for selecting the number of change-points

$$\mathbb{P}(+\,|\,x)=\frac{\sum_{a\in+}\mathcal{S}(a,x)}{\sum_{a\in+}\mathcal{S}(a,x)+\sum_{a\in-}\mathcal{S}(a,x)}, \tag{1}$$

where $+$ represents the set of positive stimuli, and $-$ the set of negative stimuli. Note that the word "set" is being used in the mathematical sense, as an unordered collection of unique elements, and thus no repetitions of items are included. The term $\mathcal{S}(a,x)$ denotes the similarity between stimuli a and x, and it is computed as an exponentially decaying function of the distance between the two stimuli:

$$\mathcal{S}(a,x)=e^{-c\cdot d(a,x)^q}, \tag{2}$$

where $d(a,x)$ is the distance between stimuli a and x, q a positive constant, and c a sensitivity parameter ($c\geq 0$). The

distance between stimuli a and x is computed by:

$$d(a,x) = \left[\sum_{i=1}^{\mathcal{D}} \omega_i \cdot |a^{(i)} - x^{(i)}|^r\right]^{\frac{1}{r}},$$

where ω_i is the attention allocated to dimension i ($\omega_i \geq 0$ and $\sum_{i=1}^{\mathcal{D}} \omega_i = 1$), r a positive constant, $a^{(i)}$ and $x^{(i)}$ the feature values of stimuli a and x on dimension i, and $\mathcal{D}$ the number of dimensions (stimuli are embedded in a $\mathcal{D}$-dimensional psychological space, in our case $\mathcal{D} = 4$). The values of the constants q and r depend on the nature of the stimuli. Since our experiments involve highly distinguishable and separable-dimension stimuli (Garner, 1974; Shepard, 1964, 1987), both constants are set equal to 1. From now on, we add the subscript "c" to the notation of the classification probability $\mathbb{P}_c$ to emphasize its dependence on the sensitivity parameter (which will vary from segment to segment in the sequel).

In this study, we examine a simplified version of the GCM where the attention given to each dimension is fixed and evenly distributed ($\omega_i = \frac{1}{\mathcal{D}}$ for each $i = 1, \ldots, \mathcal{D}$). As an initial investigation, we proposed to examine a simplified version of the GCM in order to assess the potential of the segmentation/clustering technique while reducing the complexity and computational cost of its implementation.

Although the general version of the GCM also includes additional terms, such as response-scaling and response-bias parameters, and memory-strength values, we chose not to implement these parameters in our study for two reasons. Firstly, we aimed to use the simplest possible version of the GCM. Secondly, the category concept used in the experiments had an equal number of positive and negative examples, and each block comprised a balanced representation of positive and negative stimuli. As a result, both the response-bias and memory-strength terms were unnecessary.

The likelihood of GCM on observations $z_1, \ldots, z_n$ is given by:

$$\mathcal{L}(z_1, \ldots, z_n\,;\, c) = \prod_{i=1}^{n} \left[\mathbb{P}_c\left(+ \,|\, x_i\right)\right]^{z_i} \cdot \left[\mathbb{P}_c\left(- \,|\, x_i\right)\right]^{1-z_i},$$

where n is the length of stimuli presented to a participant, x_i the i-th stimulus, z_i the classification response of stimulus x_i (1 if classified into positive stimuli, and 0 if classified into negative stimuli), and c the sensitivity parameter of GCM.

Model

Let $m \in \mathcal{M}$ be a participant, and $z_1^m, \ldots, z_{n_m}^m$ the participant's sequence of responses involving n_m data points. Also, let $Z_1^m, \ldots, Z_{n_m}^m$ be n_m random variables such that z_i^m is a realization of Z_i^m ($i = 1, \ldots, n_m$). We suppose that the process $Z_1^m, \ldots, Z_{n_m}^m$ is affected by K abrupt changes at unknown coordinates $\tau^m = \{\tau_1^m, \ldots, \tau_K^m\}$, with the convention $\tau_0^m = 1$ and $\tau_{K+1}^m = n + 1$. The K change-points define a partition of the observations into $K + 1$ segments $S_1^m, \ldots, S_{K+1}^m$ such that:

$$S_k^m = \left\{z_t^m,\, t \in \left[\tau_{k-1}^m, \tau_k^m\right)\right\}.$$

According to the segmentation/clustering model, the random variables Z_t^m follow a Bernoulli distribution of parameter the probability of classifying a stimulus x_t^m into positive stimuli (Category +) according to GCM (i.e., Equation 1):

$$Z_t^m \sim \mathcal{B}\left(\mathbb{P}_{c_k^m}\left(+ \,|\, x_t^m\right)\right), \qquad \forall t \in S_k^m,$$

where c_k^m is the sensitivity parameter of GCM associated with the segment S_k^m. The peculiarity of this method lies on the fact that the parameter c_k^m can only take P values, $c_k^m \in \{\theta_1, \ldots, \theta_P\}$. Therefore, P denotes the number of behaviors that can be assigned to segments, and $\theta_1, \ldots, \theta_P$ are the values associated with each behavior. Note that the same set of behaviors apply to all participants (within a single experiment), but not every participant needs to exhibit the entire set of behaviors. For instance, suppose participant m_1 initially classifies the stimuli randomly and then achieves a 75% correct response rate, while participant m_2 initially classifies the stimuli randomly and then achieves a perfect response rate. In this case, the method is expected to identify three distinct behaviors, even though not all participants exhibited all three.

In addition to the spatial organization of the data into segments via the partition τ^m, a secondary organization of the segments into behaviors is considered. In our context, behaviors code different learning performance (e.g., random classification, perfect classification, etc.), while the partition into segments allows the model to evolve. One can note that the parameter c_k^m is stationary on the segment S_k^m, meaning that observations on each segment are supposed independent.

The segmentation/clustering method allows GCM to adjust its performance over time through the evolution of the sensitivity parameter, without imposing constraints on how this process takes place. This unconstrained approach enables a transfer model to evolve over time while grouping subjects with similar learning progressions.

Objective

The objective of the segmentation/clustering method is to infer from observed data (i.e., participants' responses) the coordinates of the change-points as well as the values associated with each behavior. More specifically, this method aims at finding $\tau^m = \{\tau_1^m, \ldots, \tau_K^m\}$ for every participant $m \in \mathcal{M}$

and $\theta = \{\theta_1, \ldots, \theta_P\}$ such that the cost of the segmentation is minimal, given some observed data $z_1^m, \ldots, z_{n_m}^m$. Recall that $\theta_1, \ldots, \theta_P$ are common values shared by all participants within the same experiment. The smallest segmentation cost given K change-points and P behaviors parameterized by θ is as follows:

$$\mathfrak{C}_{KP}(\theta) = \sum_{m \in \mathcal{M}} \min_{\tau_1^m, \ldots, \tau_K^m} \sum_{k=1}^{K+1} \mathcal{C}_{\tau_{k-1}^m : \tau_k^m},$$

where the quantity $\mathcal{C}_{\tau_{k-1}^m : \tau_k^m}$ represents the segmentation cost of the k-th segment of participant m (i.e., S_k^m). We suppose that the segmentation cost of segment S_k^m is expressed by minus the log likelihood of GCM evaluated on S_k^m. To minimize this quantity at a fixed θ, each segment S_k^m is assigned the behavior $p_{S_k^m}$ that minimizes $-\log \mathcal{L}(S_k^m\,;\; c)$ for $c \in \{\theta_1, \ldots, \theta_P\}$. The segmentation itself is performed for each participant separately to minimize the cost. Therefore, the minimal cost of partitioning the observed data into $K+1$ segments with P behaviors parameterized by θ is the following:

$$\begin{aligned}
\mathfrak{C}_{KP}(\theta) &= \sum_{m \in \mathcal{M}} \min_{\tau_1^m, \ldots, \tau_K^m} \sum_{k=1}^{K+1} \min_{c \in \{\theta_1, \ldots, \theta_P\}} -\log \mathcal{L}(S_k^m\,;\; c) \\
&= \sum_{m \in \mathcal{M}} \min_{\tau_1^m, \ldots, \tau_K^m} \sum_{k=1}^{K+1} \min_{c \in \{\theta_1, \ldots, \theta_P\}} \sum_{j \in [\tau_{k-1}^m, \tau_k^m)} \\
&\quad (z_j - 1) \log \mathbb{P}_c\left(- \,|\, x_j\right) + \\
&\quad - z_j \log \mathbb{P}_c\left(+ \,|\, x_j\right).
\end{aligned}$$

The next step is to globally minimize $\mathfrak{C}_{KP}(\theta)$ with respect to θ, in order to find the best parameters that fit the behaviors. However, this minimization alters the value of θ. Therefore, one has to alternate between the two minimizations successively to find the global minimizer for both the segmentation and the value of the parameters. This leads to the following algorithm.

Algorithm

We used the Dynamic Programming-Expectation Maximization (DP-EM) algorithm proposed by Picard et al. (2007) to apply the segmentation/clustering model. This algorithm combines the dynamic programming (DP) algorithm used in segmentation models, with the expectation maximization (EM) algorithm used in mixture models. The principle of the DP-EM algorithm is the following: when the values associated with the P behaviors $\theta = \{\theta_1, \ldots, \theta_P\}$ are known, the coordinates of the K change-points $\tau^m = \{\tau_1^m, \ldots, \tau_K^m\}$ are computed using the DP algorithm (for each participant $m \in \mathcal{M}$), and once the coordinates of the K change-points $\tau^m = \{\tau_1^m, \ldots, \tau_K^m\}$ are estimated (for each $m \in \mathcal{M}$), the EM algorithm is used to optimize the values associated with the P behaviors $\theta = \{\theta_1, \ldots, \theta_P\}$.

The algorithm is run for a fixed number of change-points K and behaviors P. The first step consists in associating a value θ_p to each behavior $p \in \mathcal{P}$, where $\mathcal{P}$ denotes the set of P behaviors. Given $\theta = \{\theta_1, \ldots, \theta_P\}$, the second step (DP algorithm) consists in finding the coordinates of the change-points $\tau^m = \{\tau_1^m, \ldots, \tau_K^m\}$ such that the segmentation cost associated with each participant is minimal:

$$\begin{aligned}
\mathfrak{C}_{KP}^m = \min_{\tau_1^m, \ldots, \tau_K^m} \sum_{k=1}^{K+1} \min_{c \in \{\theta_1, \ldots, \theta_P\}} -\log \mathcal{L}(S_k^m\,;\; c), \\
\forall m \in \mathcal{M}
\end{aligned}$$

The third step (EM algorithm) consists in selecting among all participants the segments associated with a specific behavior p, and optimizing its value θ_p to minimize the segmentation cost of the segments associated with p:

$$\begin{aligned}
&\theta_p \in \operatorname{argmin}_c \sum_{j \text{ in a segment } S_k^m \text{ s.t. } p_{S_k^m} = p} -\log \mathcal{L}\left(z_j\,;\; c\right), \\
&\forall p \in \mathcal{P}.
\end{aligned}$$

Finally, the second and third steps are iterated multiple times to ensure convergence. Since the values θ and τ^m were stable after a few iterations, the number of iterations were set equal to 3. The algorithm is illustrated in Figure 1.

Choice of the number of behaviors

The selection of the number of behaviors was carried out by means of the adaptive method proposed by Lavielle (2005). This method aims at finding the number of behaviors $\hat{P}$ with which the log likelihood ceases to increase significantly. Let us denote

$$J_P = -\log \tilde{\mathcal{L}}_P(\hat{\tau}, \hat{\theta}) = -\max_{K=0,\ldots,K_{\max}} \left\{\log \mathcal{L}_{KP}(\hat{\tau}, \hat{\theta})\right\},$$

where $\hat{\tau}$ are the estimated coordinates of the change-points, $\hat{\theta}$ the estimated values associated with each behavior, and $\mathcal{L}_{KP}(\hat{\tau}, \hat{\theta})$ the likelihood of the model with K change-points at $\hat{\tau}$ and P behaviors. The first step consists in computing $\tilde{J}_P$ as follows:

$$\tilde{J}_P = \frac{J_{P_{\max}} - J_P}{J_{P_{\max}} - J_1} \times (P_{\max} - 1) + 1.$$

This step allows one to normalize J_P, ensuring that $\tilde{J}_1 = P_{\max}$ and $\tilde{J}_{P_{\max}} = 1$. The second step consists in computing

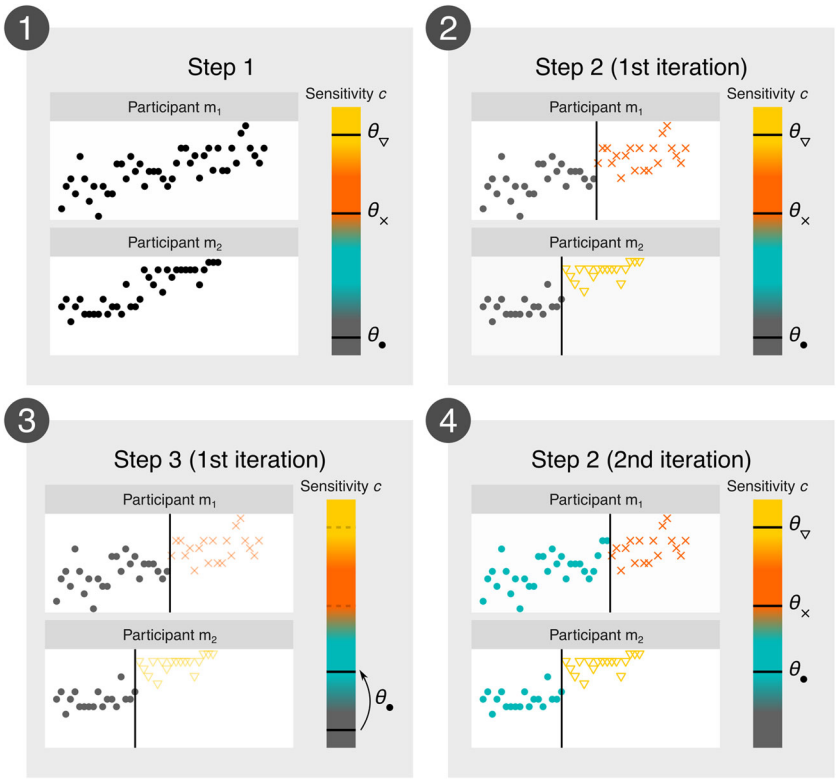

Fig. 1 Illustration of the Dynamic Programming-Expectation Maximization (DP-EM) algorithm, given three behaviors and one change-point. Data points represent the performance of participants throughout the task. The graph in the top-left corner shows the first step, where each behavior is associated with a value of the sensitivity parameter. The graph in the top-right corner shows the second step, where data points from each participant are segmented. The graph in the bottom-left corner shows the third step, where data points associated with each behavior are used to update the sensitivity parameter values of GCM. Finally, the graph in the bottom-right corner shows the second iteration of the second step, where data points from each participant are segmented using the updated sensitivity parameter values

D_P such that:

$$D_P = \tilde{J}_{P-1} - 2\tilde{J}_P + \tilde{J}_{P+1},$$

for all $P \in \{2, \ldots, P_{\max} - 1\}$. The selected number of behaviors is then given by:

$$\hat{P} = \begin{cases} \max\{P \in \{2, \ldots, P_{\max} - 1\} \text{ such that } D_P \geq \alpha\} \\ 1 \text{ if } D_P < \alpha \text{ for all } P \end{cases}$$

with α a threshold. We performed numerical simulations on classification data to tune the threshold α. In order to generate simulations that closely resemble the actual data, we simulated several instances of the GCM using the same sequence of stimuli as in Experiment 2, which is described in the subsequent section. We held the segmentation fixed and consistent for all $|\mathcal{M}| = 22$ participants. We specifically chose Experiment 2 as it involves the smallest number of participants and is therefore the most susceptible to poorly estimated parameters due to the limited amount of data. Figure 2 shows the percentage of time that the method finds the correct number of behaviors as a function of α, with different numbers of behaviors (from 1 to 4). Details about the way simulations were run are included in the caption. We found that $\alpha = 0.1$ maximizes the percentage of time that the method finds the correct number of behaviors, averaged across the selected number of behaviors P (integers ranging from 1 to 4). One can note that $\alpha = 0.1$ does not allow the method to find the correct number of behaviors when there is one behavior ($P = 1$). However, in our experiments it is reasonable to think that there are at least two behaviors (i.e., random and perfect classification) since most of the participants learned how to correctly classify the stimuli. Thus, the value $\alpha = 0.1$ suits our context.

Choice of the number of change-points

Once the number of behaviors $\hat{P}$ has been chosen, the number of change-points $\hat{K}_{\hat{P}}$ can be estimated. Let V_K be the variation of minus the log likelihood between change-points $K - 1$ and K (with $K = 1, \ldots, K_{\max}$):

$$V_K = \mathcal{L}_{K-1\hat{P}}(\hat{\tau}, \hat{\theta}) - \mathcal{L}_{K\hat{P}}(\hat{\tau}, \hat{\theta}).$$

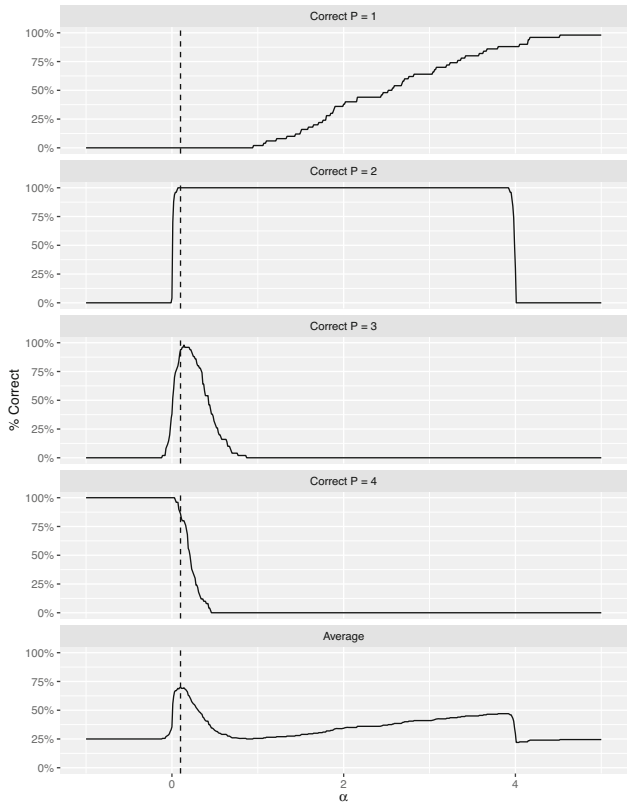

Fig. 2 Numerical simulations to tune the parameter α for selecting the number of behaviors. The graph shows the percentage of times the method identified the correct number of behaviors as a function of α, using simulated data with different numbers of behaviors P (integers ranging from 1 to 4). Each number of behavior was simulated 50 times. The values of the sensitivity parameter c associated with each behavior were as follows: 15 for $P = 1$; 0.005 and 30 for $P = 2$; 0.005, 13, and 30 for $P = 3$; and 0.005, 7, 14, and 30 for $P = 4$. These specific values were selected to enhance the distinguishability of the behaviors (see Figure 11 for an illustration of the subtle relationship between the sensitivity parameter and the accuracy rate). The same sequence of stimuli used in Experiment 2 was used to run the simulations. Data was generated such that each participant was affected by a single change-point ($K = 1$), equidistant from the participant's first and last stimuli. We set $P_{\max} = K_{\max} = 5$. The graph on the bottom shows the percentage of correct response, averaged across the previous graphs (P ranging from 1 to 4). Dashed lines indicate the selected value for α

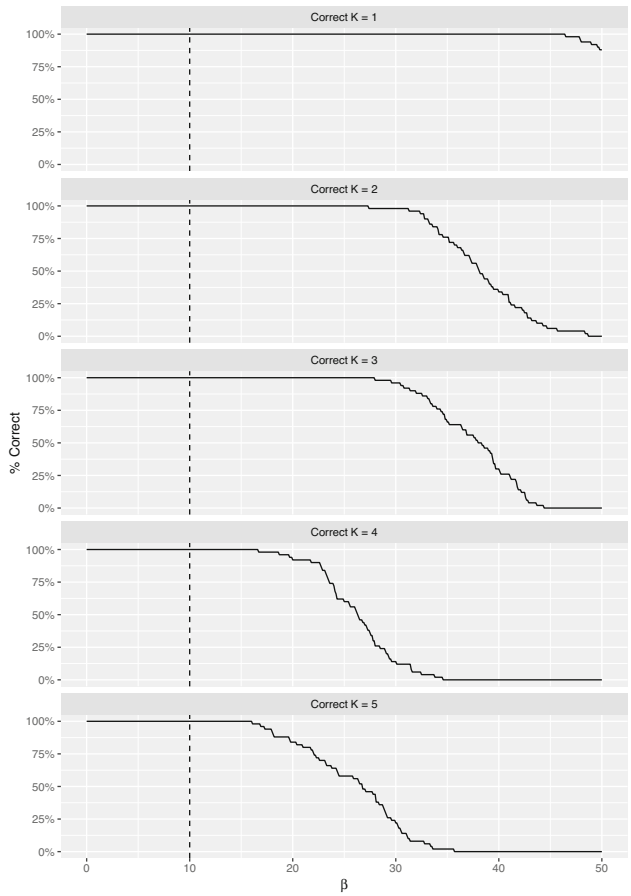

Fig. 3 Numerical simulations to tune the parameter β for selecting the number of change-points. The graph shows the percentage of times the method identified the correct number of change-points as a function of β, using simulated data with different numbers of change-points K (integers ranging from 1 to 5). Each number of change-points was simulated 50 times. Data was simulated using two behaviors ($P = 2$), and the sensitivity parameter values associated with these two behaviors were 0.005 and 15. These specific values were selected to enhance the distinguishability of the behaviors (see Figure 11 for an illustration of the subtle relationship between the sensitivity parameter and the accuracy rate). The same sequence of stimuli used in Experiment 2 was used to run the simulations. Data was generated such that each participant was affected by K change-points (integers ranging from 1 to 5), equidistant from the participant's first and last stimuli. We set $K_{\max} = 10$. Dashed lines indicate the selected value for β

The selection of the number of segments is given by:

$$\hat{K}_{\hat{P}} = \begin{cases} \max\{K \in \{1, \ldots, K_{\max}\} \text{ such that } V_K \geq \bar{V}\} \\ 0 \text{ if } \sigma_V < \beta \end{cases} \quad (3)$$

where $\bar{V}$ is the mean of $\{V_K,\ K = 1, \ldots, K_{max}\}$, σ_V its standard deviation, and $\beta \geq 0$ a tuning parameter. This method allows one to find the highest number of change-points associated with a significant decrease in minus the log likelihood. We chose to implement a method based on slope heuristics because it has been shown to outperform other methods, such as AIC and BIC, in terms of accuracy and reliability, and is more robust to deviations from the model assumptions (Picard et al., 2007). We conducted numerical simulations to both tune the parameter β and assess the efficacy of the method. Again, to generate artificial data that closely resemble the actual data, we used the same sequence of stimuli as in Experiment 2. Figure 3 shows the percentage of time that the method finds the correct number of change-points as a function of β, with simulated data having different number of change-points K (from 1 to 5). Details about the way simulations were run are included in the caption. One can observe that the method provides a correct answer in every case and simulation, for any $\beta \leq 15$. We decided to take $\beta = 10$.

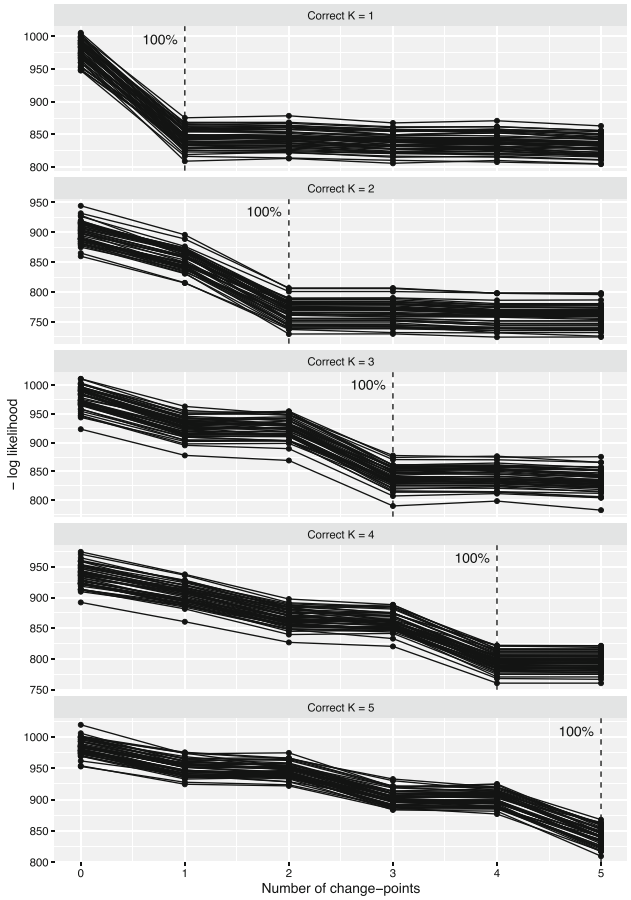

Fig. 4 Numerical simulations when $\beta = 10$. The graph shows minus the log likelihood as a function of the number of change-points, on simulated data with different numbers of change-points K (integers ranging from 1 to 5). For each number of change-points, we run 50 simulations. In each one of the 50 simulations, the method found the same result indicated with dash lines. This means that the method (with $\beta = 10$) found the correct number of change-points 100% of the time, in each case

Figure 4 shows minus the log likelihood as a function of the number of change-points, with simulated data having different number of change-points K (integers ranging from 1 to 5) and $\beta = 10$. The percentage of time the method found the correct number of change-points is included in the graph. The method allowed us to find the correct number of change-points in every simulation (amounting to 50 simulations per selected number of change-points).

Numerical simulations

Classification data are particularly complex. Indeed, predictions of GCM (and of categorization models in general) are expressed in terms of probability, while classification data are expressed in terms of binary responses (1 when participants classified stimuli into positive stimuli, and 0 otherwise). Therefore, intrinsic noise within the data can be very high.

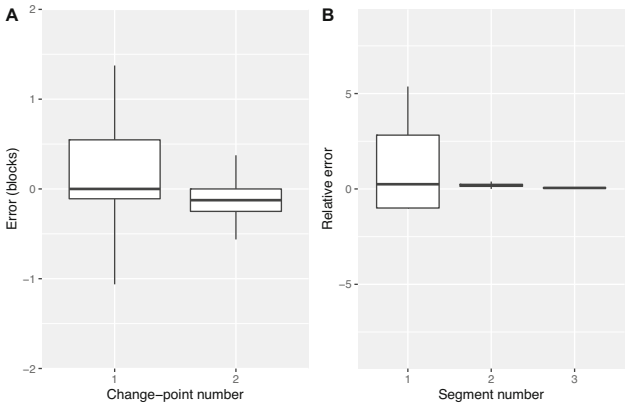

Fig. 5 Error of the segmentation/clustering technique in detecting the coordinate of the change-points (Figure A) and the value of the sensitivity parameter c (Figure B). In A, the error is expressed in terms of number of blocks. In B, the relative error is defined as $\frac{c-\hat{c}}{c}$. The number of behaviors P was set equal to 3, the number of change-points K was set equal to 2, and the values of the sensitivity parameter c were set equal to 0.2, 7, and 19 for the first, second, and third segment, respectively. These specific values for the sensitivity parameter were selected to enhance the distinguishability of the behaviors (see Figure 11 for an illustration of the subtle relationship between the sensitivity parameter and the accuracy rate). The same sequence of stimuli used in Experiment 2 was used to run the simulations. Data was generated such that the change-points affecting participants' progression were equidistant from each participant's first and last stimuli

Numerical simulations were conducted to assess the reliability of the segmentation/clustering technique on classification data, specifically.

Simulations were run with a fixed number of change-points K and behaviors P. Figure 5 shows the error of the segmentation/clustering technique in detecting the coordinates of the change-points (Figure A) and the value of the sensitivity parameter c (Figure B), when $P = 3$, $K = 2$, and $c = 0.2$, 7, and 19 for the first, second, and third segment, respectively. In Figure 5A, the coordinates of the two change-points were estimated with a high accuracy (± 1 block for the first change-point, and $\pm$ half a block for the second change-point). Here, the term "block" refers to a full cycle of stimuli to be classified. For instance, if participants are required to classify 16 stimuli (as in our experiments), then each block would consist of 16 stimuli. Further details on the experiments and the nature of the stimuli are given in the next section.

In Figure 5B, the lowest value of the sensitivity parameter c were estimated with a medium accuracy, while the highest values were estimated with a high accuracy. This is not surprising since the predictions of the model are in the surrounding of 0.5 when the sensitivity parameter is close to 0, increasing the noise within the data. Note that values of c equal to or greater than 20 result in perfect classification (when the number of blocks is small). Since participants met the learning criterion after the successfully completion of 4

blocks, the difference between $c = 20$ and $c > 20$ would not be noticeable in a such a short time window. Therefore, we limited the sensitivity parameter to be smaller than 20. Simulations with different number of behaviors and change-points gave similar results.

Experiments

In the current section, we describe three novel experiments, in which presentation order was manipulated. The experiments were performed in accordance with relevant guidelines and informed consent was obtained from all participants prior to participation.

Experiment 1

This experiment is closely based on that of Mathy and Feldman (2009), modified in several ways to keep consistency across all three experiments. In particular, while Mathy and Feldman (2009) only manipulated the order of positive stimuli (i.e., members of Category +), here we control the order of both positive and negative stimuli (i.e., members of both categories). In addition, in the present study only one concept was administered to participants instead of two concepts as in Mathy and Feldman (2009).

Participants

The participants were 68 freshmen and sophomores at the Université de Franche-Comté (France), who received course credit in exchange for their participation.

Choice of concept studied

Each participant was administered a concept defined over four Boolean dimensions. According to the classification of Feldman (2003), this concept is called $12_{4[8]}$ (Figure 6, on the top) to indicate that it is the 12^{th} in a set of 4-dimensional concepts consisting of 8 positive stimuli. The choice of a four-dimensional concept is justified by the fact that the number of items to be classified ($2^4 = 16$) is large enough to allow the detection of effects of presentation order, but small enough to allow its memorization.

This concept has interesting properties: *i)* it is moderately complex, and *ii)* it is characterized by a substructure made of several well-defined subcategories. This substructure made of subcategories is more detectable by considering the compressed formula of the studied concept $12_{4[8]} \cong a'(bc)' + ad'(bc' + b'c)$. We use here a standard notation (Feldman, 2000, 2003), in which a' refers to negation ($\neg$) of feature a (a and a' are the two dimension values that can be taken by dimension A), ab refers to the conjunction ($\wedge$) of

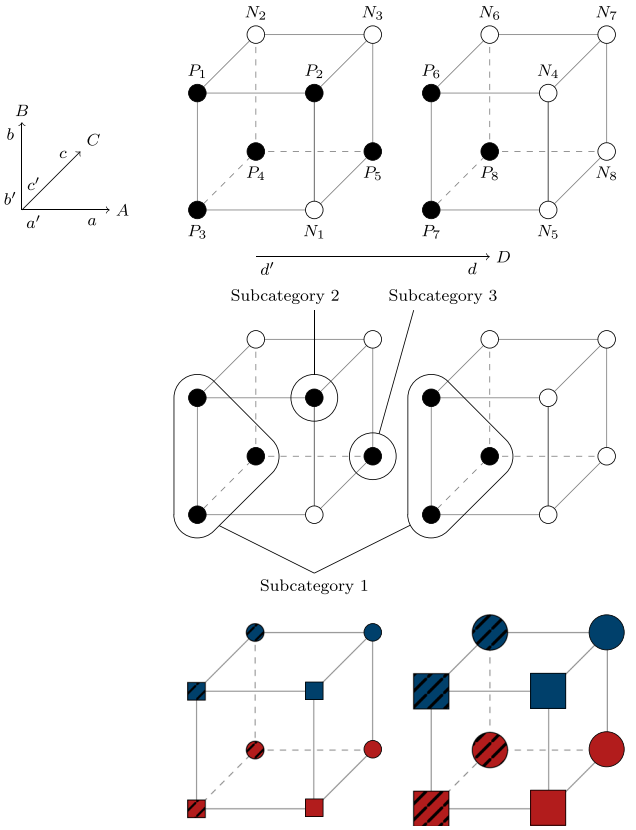

Fig. 6 Illustration of the concept and stimulus items of Experiment 1. On the top, the concept $12_{4[8]}$ according to Feldman's classification (Feldman, 2003). Positive stimuli are indicated with black circles, while negative stimuli are indicated with white circles. The notation $12_{4[8]}$ refers to the fact that this concept is the 12^{th} in the Feldman's list of 4-dimensional concepts consisting of 8 positive stimuli. In the middle, the substructure made of subcategories in concept $12_{4[8]}$. To avoid overburden the figure, only subcategories of the positive stimuli are shown. On the bottom, an example of the $2^4 = 16$ stimulus items presented to participants. The items varied along four Boolean dimensions (Shape, Color, Size, and Filling pattern). To make the figure more readable, we illustrated plain and striped items instead of cross-hatched and striped items

a and b, and $a + b$ to their disjunction ($\vee$). The $\cong$ symbol indicates that any other concept isomorphic to this formula can be labelled $12_{4[8]}$. A verbal example of the compressed formula using the illustration in the bottom panel of Figure 6 is "all striped objects except blue circles, as well as the small blue plain square and the small red plain circle".

The substructure made of subcategories in concept $12_{4[8]}$ is represented in Figure 6 in the middle (the figure only shows the subcategories of the positive stimuli). Subcategory 1 represents six of the eight members of the concept (P_1, P_3, P_4 P_6, P_7 and P_8), corresponding to the first disjunctive clause $a'(bc)'$ in the compressed formula. These six items can collectively be represented by a verbal expression such as "all a' except bc". By contrast, Subcategories 2 and 3 consist of one object each (P_2 for Subcategory 2 and P_5 for Subcategory 3) and correspond to the expansion

of the second clause in the compressed formula ($abc'd'$ and $ab'cd'$, respectively). Therefore, Subcategory 1 plays the role of a salient "rule", while Subcategories 2 and 3 play the role of "exceptions". Following Mathy and Feldman (2009), we hypothesize that grouping $12_{4[8]}$ into these particular subcategories is beneficial to learning.

Stimuli

Stimulus items varied along four Boolean dimensions (Shape, Color, Size, and Filling pattern). Rotation and permutation were randomized for each participant, meaning that dimension A could correspond to Shape, or Color, or Size, or Filling pattern depending on the participant, and that features within dimensions were randomly drawn and permuted (for instance, $a' = blue$ and $a = red$, or $a' = red$ and $a = blue$, or $a' = green$ and $a = red$, etc.). The choice of two values for each feature was randomly chosen among these features: triangle, square, or circle for Shape; blue, pink, red, or green for Color; small or big for Size; and hatched or cross-hatched for Filling pattern. Overall, the combination of these four separable dimensions (Garner, 1974) formed 16 single unified items (e.g., a small hatched red square, a big cross-hatched blue circle, etc.).

Ordering of stimuli

We used two types of presentation orders: a rule-based order and a similarity-based order. These were the orders that best facilitated learning in Mathy and Feldman's study (2009). Presentation order was a between-subject manipulation. One type of presentation order was randomly chosen for a given participant beforehand and then applied across the blocks. Here, a block is defined as a full cycle of 16 items (8 positive and 8 negative). Negative stimuli were randomly intermingled with positive stimuli and a variable presentation across blocks was used, meaning that each new block (although constrained to a given order type) was newly randomized. Because categories were randomly alternated and a variable presentation across blocks was used, we refer to the context of this experiment as Random-Variable.

Unlike the study by Mathy and Feldman (2009), the negative stimuli were also grouped into subcategories. Subcategory 2 within the negative stimuli was defined by the negation of $(bc)'$ on the a' feature (i.e., $a'bc$) and included items N_2 and N_6. Subcategory 1 within the negative stimuli was defined by the negation of $d'(bc' + b'c)$ on the a feature (i.e., $a(d'(bc' + b'c))'$) and included the rest of the negative items (N_1, N_3, N_4, N_5, N_7, and N_8). Therefore, the negative subcategories were simply regarded as an inversion of the positive subcategories.

In the rule-based order, the positive items were randomly drawn from Subcategory 1 until all 6 stimuli were presented. Likewise for the negative items belonging to Subcategory 1. These were followed by the positive items in Subcategory 2 and Subcategory 3 (the item in Subcategory 2 was presented strictly before the item in Subcategory 3), and by the negative items belonging to Subcategory 2 (in random order). Thus in the rule-based order, all members of the largest subcategory were presented first (in random order) and separated from exceptional members, in order to promote the abstraction of the simplest rules by participants. The presentation within subcategories was randomized to facilitate the extraction of the relevant features.

In the similarity-based order, the first item was randomly selected and subsequent items were randomly chosen from those maximally similar to the previous item until the set of stimuli was exhausted. The negative stimuli were also similarity-based ordered and ties were resolved randomly. Similarity was computed on a trial-by-trial basis so as to maximize inter-item similarity locally, a method which did not guarantee a maximized inter-item similarity over an entire block, but which offered a greater number of possible orders. Similarity between two stimuli x and y was computed using:

$$s_{xy} = \sum_{i=1}^{\mathcal{D}} \mathbb{1}_{\{x_i = y_i\}},$$

which allows the count of the common features shared by the two stimuli. In the above formula, x_i and y_i are the feature values of stimuli x and y on dimension i and $\mathcal{D}$ represents the dimension of the space in which items are embedded (which is four in our experiment). Be careful not to confuse the definition of similarity used to order the stimuli in the similarity-based order with the definition used in the GCM (Equation 2). The most important aspect of this procedure is that the ordering does not necessarily respect the subcategory boundaries targeted in the rule-based order, as similarity steps can cross in and out of subcategories. For instance, the stimulus P_1 can be followed by stimulus P_2.

Procedure

There was no warm-up session (such as learning a simple one-dimensional concept) so that participants would not think that the task consisted in searching for simplistic rules. However, participants were briefly instructed before the task began. Each participant was asked to learn a single $12_{4[8]}$ concept following either a rule-based order or a similarity-based order (half of the participants were assigned to the rule-based order).

The task was computer-driven and participants were tested individually during a one-hour single session (including briefing and debriefing). Participants sat approximately 60 cm from a computer on which stimulus items were presented

one at a time in the upper part of the screen. They learned to sort the stimulus items using two keys, and successful learning was encouraged by means of a progress bar. The positive and negative categories were associated with the up and down keys respectively, and by two category pictures on the right hand side of the screen. A virtual frame for the categories faced the frame that encompassed the stimulus on its left. The frame for the categories displayed a schoolbag at the top, and a trash can at the bottom (to match the response keys). Each time a response key was pressed, the corresponding picture was displayed for two seconds along with feedback, while the opposite picture was hidden for two seconds. After each response, feedback indicating a correct or incorrect classification was given at the bottom of the screen for two seconds. The two category pictures reappeared whenever a new stimulus was presented.

The participants scored one point for each correct response which was shown on the progress bar. To regulate the learning process, each response had to be given in less than eight seconds (resulting in a maximum of 10 seconds between two stimuli when the participants got a 'Too late' message that lasted two seconds). If the response was given too late, the participants would lose three points on the progress bar. This was thought to prevent the participants from skipping the most difficult stimuli without any penalty. The number of empty boxes in the progress bar was $4 \times 16 = 64$. One empty box was filled whenever a correct response was given, but the progress bar was reset in case of an incorrect response. This criterion was identical to the one used by Shepard et al. (1961) in their first experiment and by Mathy and Feldman (2009). As a consequence, successful completion of the experiment required participants to accurately classify stimuli on four consecutive blocks consisting of 16 stimuli each (64 stimuli in total). This stringent criterion required participants to correctly classify all stimuli, including the exceptions, thereby preventing them from resorting to partial solutions. There was no limit on the number of learning blocks, and participants were free to withdraw from the study at any time.

Experiment 2

Experiment 2 was designed to investigate the effect of a constant presentation across blocks. By "constant presentation" we refer to the use of the same predetermined sequence of 16 stimuli across all blocks. We hypothesized that such an order could facilitate both the perception of commonalities within categories (when two stimuli of the same category are presented repeatedly and contiguously) and the perception of contrasts between categories (when two stimuli of different categories are presented repeatedly and contiguously), which would lead to form an abstraction. From an exemplar point of view, constant orders were thought to limit the number of temporal associations between stimuli, which should therefore reinforce the limited set of associations between the memory traces. In this experiment, we used the same concept, stimuli, and types of orders as in Experiment 1.

Participants

The participants were 22 freshmen and sophomores at the Université de Franche-Comté (France), who received course credit in exchange for their participation.

Procedure

The procedure was similar to Experiment 1, except that a constant presentation across blocks was used. This procedure can presumably help participants perceive sub-patterns of responses (e.g., $-+++$) that can be used to classify instances blindly. For instance, after noticing that a $+++$ patterns occurs after a "large red hatched square", this pattern can be used as a cue to correctly classify three instances in a row without paying attention to the stimuli. This is the reason why this condition was tested with a small sample of participants. The number of participants per type of presentation order was balanced. Because categories were randomly alternated and a constant presentation across blocks was used, we refer to the context of this experiment as Random-Constant.

Experiment 3

This experiment explores the effect of blocking negative and positive stimuli. In Mathy and Feldman (2009) and in earlier studies (Elio and Anderson, 1981, 1984), negative stimuli were interleaved with positive stimuli in order to emulate an ordinary random presentation. More recently, Mezzadri et al. (2022b) have investigated interactions between various order manipulations including interleaving vs. blocking (in which positives and negatives are segregated) with a relatively simple concept. Here, we explore the effect of interleaved vs. blocked presentation using a more complex concept. One hypothesis is that when stimuli are blocked the perception of the commonalities within categories is favored in the rule-based order, hence resulting in faster learning. However, the perception of contrasts between categories might be enhanced because of the immediate juxtaposition of positives and negatives. It is thus difficult to decide between the opposite effects of blocked vs. interleaved presentations without knowing exactly the type of category being studied (Carvalho and Goldstone, 2015). Here, because we use a difficult concept with highly discriminable categories (in which the stimuli are dissimilar both within and between categories), a blocked presentation should result in better performance using a rule-based presentation. Again, the same concept, stimuli, and types of orders as in Experiment 1 were used.

Participants

The participants were 46 freshmen and sophomores at the Université de Franche-Comté (France), who received course credit in exchange for their participation.

Procedure

The procedure followed that of Experiment 2 with the exception of two modifications: *i)* the use of a fully-blocked presentation, where positive stimuli were always presented before negative stimuli (Clapper and Bower, 1994, 2002); and *ii)* the introduction of random blocks. Random blocks were included to prevent participants from guessing the correct responses without attending to the stimuli (i.e., the participants could categorize the stimuli blindly just by pressing eight times on the same category). Each fully-blocked block was followed by a random block, with a 5-second pause in between them. In the random blocks, stimuli were presented randomly within each block, with the sequence of 16 stimuli varying from block to block. Participants received feedback after each trial. We refer to the context of this experiment as Blocked-Constant. Again, half of the participants were assigned to the rule-based order.

Table 2 Mean inter-item similarity (Table A) and number of participants who did not complete the task (Table B), as a function of presentation order and experiment. The maximal inter-item similarity is 3 in all experiment, since two contiguous four-dimensional items cannot have more than 3 features in common. The term "unsuccessful" participants refers to those individuals who did not meet the learning criterion

A

	Rule-based	Sim.-based
Experiment 1	1.89	2.23
Experiment 2	1.86	2.25
Experiment 3	2.25	2.36

B

	Rule-based	Sim.-based
Experiment 1		
Successful	21	17
Unsuccessful	13	17
Experiment 2		
Successful	11	9
Unsuccessful	0	2
Experiment 3		
Successful	20	16
Unsuccessful	3	7

Data Analysis

The mean inter-item similarity for each presentation order and experiment is given in Table 2A. As expected, the average inter-item similarity was higher for the similarity-based order than for the rule-based order in all experiments. One can note that in Experiment 3 blocking greatly increased the mean inter-item similarity. Figure 7 (on the top) shows the average number of blocks that participants took to meet the learning criterion, as a function of presentation order and experiment (graphs showing all participants were plotted separately from the ones with successful participants alone). One can note that on average in Experiment 2 and Experiment 3 participants in the rule-based condition completed the task faster than participants in the similarity-based condition.

Two survival analysis techniques were performed to study the influence of presentation order (rule-based vs. similarity-based) on the time required by participants to complete the task: the Kaplan-Meier survival curves and the Cox proportional-hazards model. Survival analysis techniques were preferred to other analyses because of their ability to take into account participants who did not meet the learning criterion. The number and presentation order of participants who did not complete the task are shown in Table 2B. None of the participants were removed from the analyses. Although the participants' learning progression was not statistically analyzed, we include a graph (Figure 7, on the bottom) that shows the average percentage of correct responses among participants within a same condition as a function of block number over the course of the experiments. Again, a faster progression in the rule-based condition as compared to the similarity-based condition can be observed in Experiment 2 and Experiment 3.

Kaplan–Meier survival curves

The Kaplan-Meier estimator (Kaplan and Meier, 1958) allows one to estimate the expected duration of time until an event of interest occurs. Our event of interest is the time at which participants met the learning criterion. Figure 8 shows the survival probability for each type of presentation order, as a function of block number and experiment. The survival probability shows how participants assigned to a given condition are likely to continue the task (and consequently, to not meet the learning criterion). A log-rank test was performed to evaluate the difference between survival curves. The log-rank test was significant in Experiment 2 and Experiment 3 (p-value $= 0.0051$ in Experiment 2, and p-value $= 0.04$ in Experiment 3). This shows that learning was faster in the rule-based order as compared to the similarity-based order in both Experiment 2 and Experiment 3. Despite Experiment 1

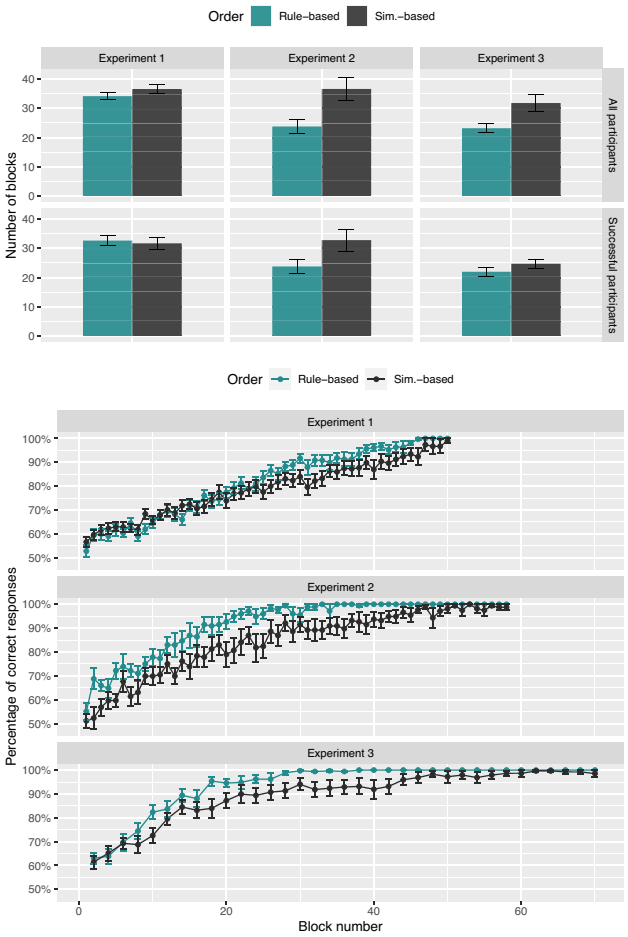

Fig. 7 Participants' learning time and progression as a function of presentation order in Experiments 1-3. On the top, average number of blocks taken by participants to meet the learning criterion, as a function of presentation order. Graphs showing all participants were plotted separately from the ones with successful participants alone. The term "successful" participants refers to those individuals who met the learning criterion. On the bottom, average percentage of correct responses among participants within the same presentation order, as a function of block number. In Experiment 3, only performance across random blocks are plotted

closely resembling the original study by Mathy and Feldman (2009), we did not observe any advantage of the rule-based order over the similarity-based order. This lack of advantage may be attributed to the manipulation of both positive and negative examples in the present study. In other words, the manipulation of both positive and negative examples may have interfered with both types of order, resulting in similar performance levels.

Cox proportional-hazards model

The Cox model (Cox, 1972) is a survival analysis technique that allows one to simultaneously account for multiple variables. Therefore, this analysis allows us to additionally examine the impact of context (Random-Variable in Experiment 1 vs. Random-Constant in Experiment 2 vs. Blocked-Constant in Experiment 3) on learning speed, while investigating our main manipulation of interest (rule-based vs. similarity-based). Figure 9 illustrates the result of the Cox model as a function of presentation order and context. The graph shows that contexts Random-Constant and Blocked-Constant increased participants' hazard ratio as compared to the reference condition (i.e., context Random-Variable). This means that these contexts were found to help participants to meet the learning criterion faster. The impact of both contexts was found significant (p-value < 0.001 for both the Random-Constant and Blocked-Constant contexts). By contrast, the similarity-based order reduced participants' hazard ratio as compared to the rule-based order. This impact was found significant (p-value < 0.001), showing that learning was slower in the similarity-based condition.

Results

Here, we present the results of the application of the segmentation/clustering technique to Experiments 1-3 using GCM as underlying model. The method for selecting the number of behaviors with $\alpha = 0.1$ found 4 behaviors in Experiment 1 and 2, and 3 behaviors in Experiment 3 (see Figure 10, top). The method for selecting the number of change-points with $\beta = 10$ found 1 change-point for all three experiments (see Figure 10, bottom). An evaluation of the fit of GCM when the segmentation/clustering technique is applied to the data can be found in Supplementary material B.

Table 3A displays the values of the sensitivity parameter c associated with each behavior of Experiments 1-3. To facilitate the comprehension of the results, Figure 11 shows the impact of varying sensitivity parameter values on the expected accuracy range. For instance, for $c = 0$ the average accuracy rate is 50%, whereas for $c = 10$ it is 85%. Generally, higher values of the sensitivity parameter correspond to a greater proportion of correct responses per block. Because values of c equal to or greater than 20 resulted in perfect classification with a small number of blocks (as previously mentioned), the upper bound of the sensitivity parameter was set equal to 20. In Experiments 1 and 2 there were 4 learning regimes (low, medium, high, and perfect/almost perfect classification), whereas in Experiment 3 there were 3 learning regimes (low, high, and perfect/almost perfect classification).

Figure 12 (on the top) shows the result of the application of the segmentation/clustering technique with 4 behaviors and 1 change-point to 3 participants of Experiment 2. The results on the remaining participants of Experiment 2, as well as those of Experiments 1 and 3 are shown in Supplementary material C. Figure 12 (on the bottom) shows the density func-

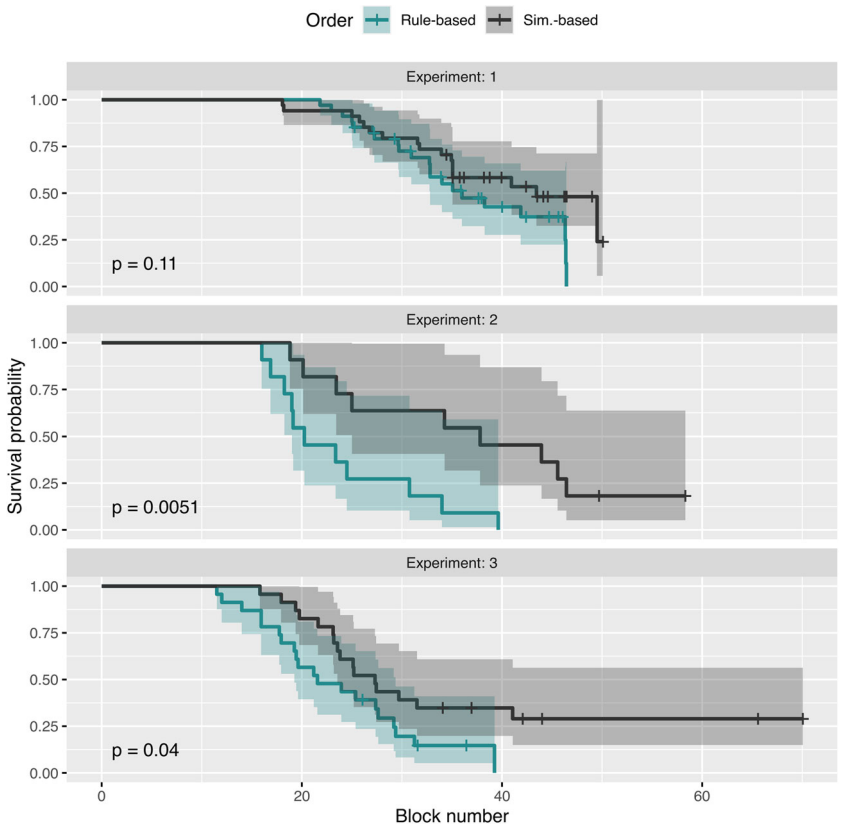

Fig. 8 Kaplan-Meier survival curves for each presentation order as a function of block number in Experiments 1-3. Transparent areas represent the 95% confidence intervals. p-values of the log-rang test assessing the difference between survival curves of participants in the rule-based and similarity-based orders are showed on the bottom-left side of each graph

tion of the coordinate of the change-point (in terms of block number) in Experiments 1-3. One can observe that change-points in Experiment 1 have a higher coordinate than those in Experiments 2 and 3 (the average coordinate is 359, 259, and 228 stimuli in Experiments 1-3, respectively). The two-sided Wilcoxon-Mann-Whitney test shows that the difference was significant (p-value $= .006$ between Experiments 1 and 2, and p-value $< .001$ between Experiments 1 and 3). This is

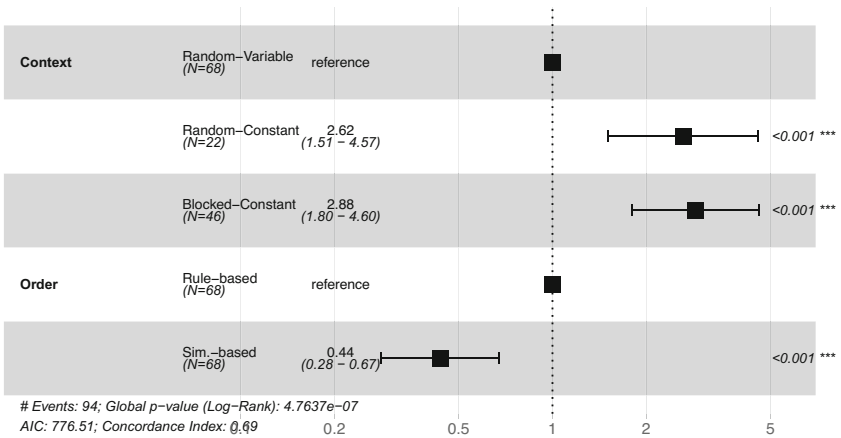

Fig. 9 Results of the application of the Cox model as a function of presentation order and context. In this analysis, Experiments 1-3 have been aggregated together to determine the effect of the three contexts. Hazard ratios and their 95% confidence intervals are showed for each condition in the middle of the graph. Statistical significance of the Wald test is showed for each condition on the right side of the graph

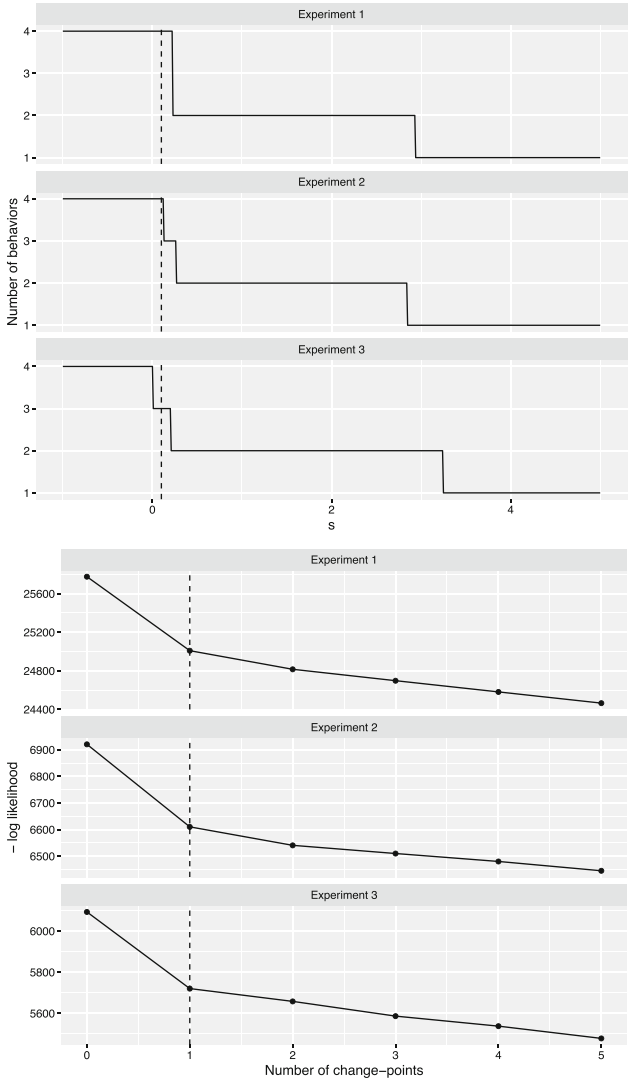

Fig. 10 Chosen number of behaviors (top) and change-points (bottom), for Experiments 1-3. On the top, number of behaviors as a function of α. The value $\alpha = 0.1$ (i.e., the dashed lines) determines the number of chosen behaviors. On the bottom, minus the log likelihood as a function of the number of change-points. Dashed lines indicate the number of chosen change-points determined by Equation 3 with $\beta = 10$. The values for α and β were found through numerical simulations (see Sections 2.5 and 2.6)

Table 3 Values of the sensitivity parameter c associated with each behavior in Experiments 1-3 (Table A); and number of participants N and average coordinate of the change-point T (expressed in terms of number of stimuli) as a function of presentation order and learning path in Experiments 1-3 (Table B). By learning path we mean the sequence of behaviors exhibited by participants. We remind that 4 behaviors were found in Experiments 1 and 2, and 3 behaviors in Experiment 3

A				
	Behavior 1	Behavior 2	Behavior 3	Behavior 4
Experiment 1	3.6	7.4	13	20
Experiment 2	4.2	9.4	15.2	20
Experiment 3	5.1	13.3	20	-
B				
Learning path	Rule-based		Sim.-based	
Experiment 1	*N*	*T*	*N*	*T*
1	1	-	0	-
1-2	7	299	10	396
1-3	15	302	12	346
1-4	2	496	2	716
2-1	1	12	0	-
2-3	2	344	5	435
2-4	6	372	5	426
Experiment 2				
1-2	1	248	5	315
1-3	5	184	4	257
1-4	2	249	1	337
2-4	3	295	1	200
Experiment 3				
1	1	-	0	-
1-2	4	117	9	173
1-3	18	88	13	125
3-1	0	-	1	2

coherent with the finding that participants in Experiments 2 and 3 met the learning criterion sooner than participants in Experiment 1 (see Section 4).

Two analyses were performed to determine whether the influence of presentation order can also be found in segmentation/clustering results. The first analysis investigated a potential relation between learning path and presentation order, where by learning path we mean the sequence of behaviors exhibited by participants. Table 3B shows the number of participants N as a function of presentation order and learning path in Experiments 1-3. One can observe that participants having a high or perfect/almost perfect ending regime (i.e., 3 or 4) in Experiments 2 and 3 are more in the rule-based order than in the similarity-based order. Inversely, participants having a medium ending regime (i.e., 2) are more in the similarity-based order than in the rule-based order. For instance, in Experiment 3 the number of participants having a learning path 1-3 is 18 in the rule-based order vs. 13 in the similarity-based order. However, the Fisher's exact tests at a 5%-level were not significant (p-value $= .71$ in Experiment 1, p-value $= .31$ in Experiment 2, and p-value $= .14$ in Experiment 3). This is probably due to the limited number of participants.

The second analysis examined the coordinate of the change-points, as a function of presentation order and learning path. Participants whose performance worsened over time showing a negative learning path (amounting to 2 in Experiment 1, and 2 in Experiment 3) were removed from the analysis. Table 3B shows the average coordinate of the

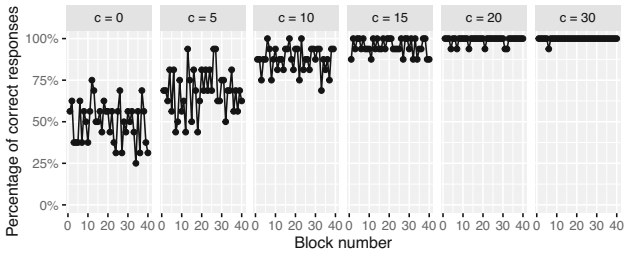

Fig. 11 Impact of varying the value of the sensitivity parameter c on the accuracy rate

change-points (expressed in terms of number of stimuli, and denoted by T), as a function of presentation order and learning path in Experiments 1-3. One can observe that the average coordinate of the change-points is higher in participants in the similarity-based order than in those in the rule-based order (except for learning path 2-4 in Experiment 2). A one-sided Wilcoxon-Mann-Whitney test at a 5%-level was conducted

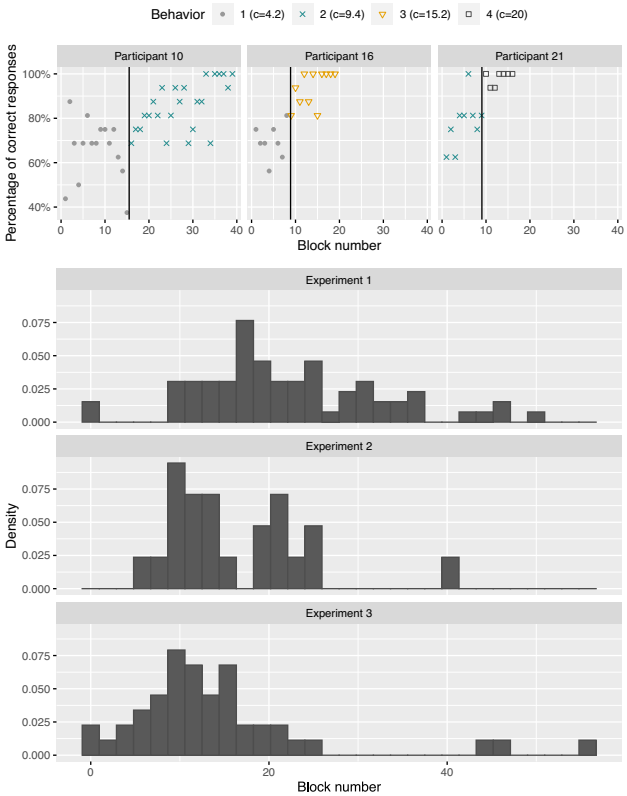

Fig. 12 Results of the application of the segmentation/clustering technique. On the top, visualization of the segmentation/clustering method with $P = 4$ behaviors and $K = 1$ change-point to Experiment 2. Data points represent the performance of participants throughout the task. To make the graph more readable, only 3 participants (Participants 10, 16, and 21) among the 22 were selected. The remaining participants of Experiment 2, as well as those of Experiments 1 and 3 are shown in Supplementary material C. On the bottom, density function of the coordinate of the change-point (in terms of block number) in Experiments 1-3

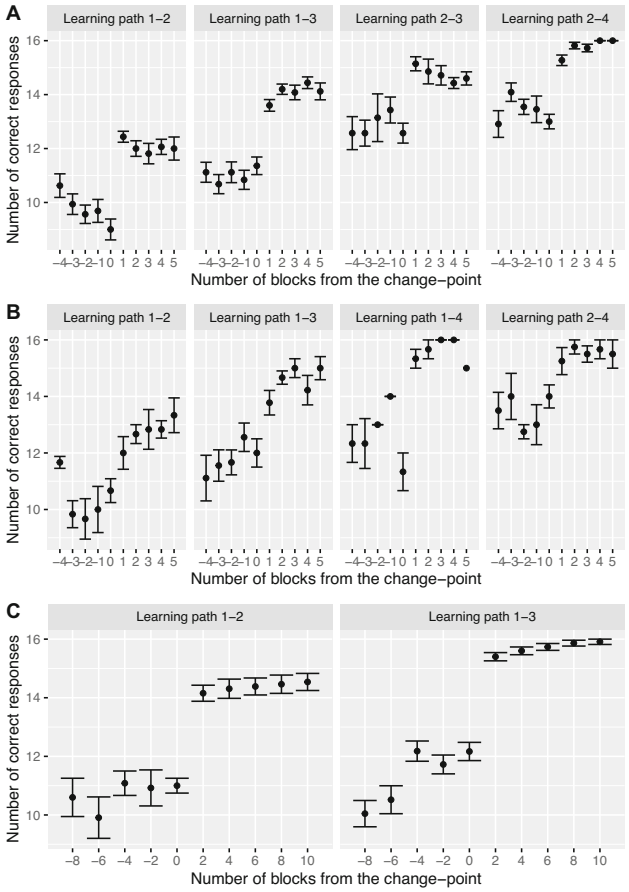

Fig. 13 Backward learning curves for Experiments 1-3 (graphs A-C, respectively). Each graph shows the average number of correct responses from subjects within the same learning path as a function of the distance from the change-point. The numbering of the learning paths indicates the sequence of behaviors showed by participants. For instance, by "Learning path 1-2" we refer to the subjects who showed behavior 1 followed by behavior 2. In Experiment 3 (graph C), only random blocks were included. Learning paths with fewer than two subjects or with less than two backward/forward performance scores per behavior were excluded

to assess whether this difference is significant. The test was only significant in Experiment 3 (p-value $= .066$ in Experiment 1, p-value $= .19$ in Experiment 2, and p-value $= .049$ in Experiment 3). Although the test was not significant in Experiment 2 as found in our previous analysis, the segmentation/clustering technique still captured main tendencies within the data.

Transitions between behaviors can be more effectively identified through a backward learning curve (Smith et al., 2004; Hayes, 1953; Zeaman and House, 1963; Smith et al., 1993). This curve is obtained by aligning the change-point coordinates of subjects within the same learning path and then plotting performance backward and forward from that relative time. Figure 13 shows the backward learning curves for Experiments 1-3. In Experiments 1 and 3, participants

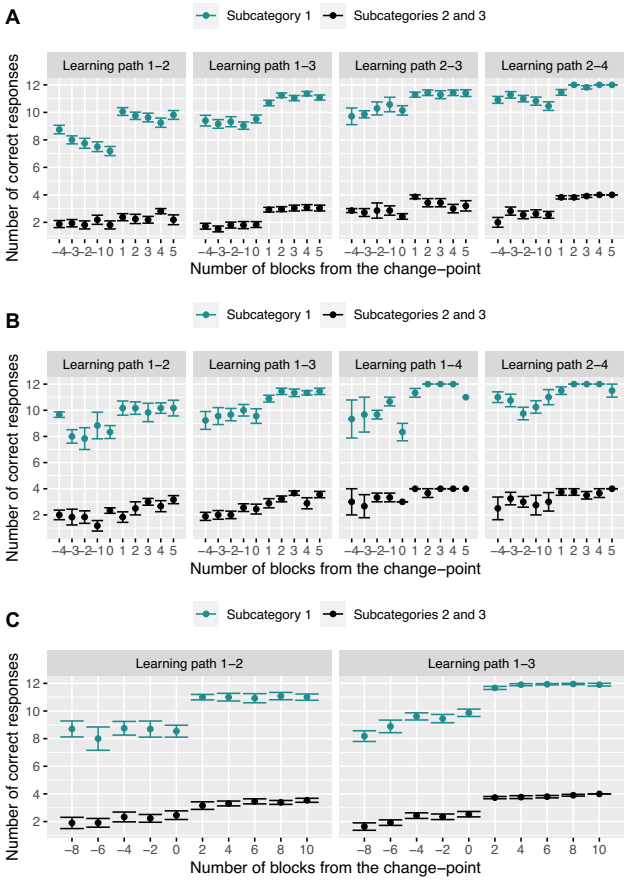

Fig. 14 Backward learning curves for Experiments 1-3 (graphs A-C, respectively), with items belonging to different subcategories plotted separately. Each graph shows the average number of correct responses from subjects within the same learning path as a function of the distance from the change-point. The items belonging to the rule (Subcategory 1) and exceptions (Subcategories 2 and 3) are plotted separately. The numbering of the learning paths indicates the sequence of behaviors showed by participants. For instance, by "Learning path 1-2" we refer to the subjects who showed behavior 1 followed by behavior 2. In Experiment 3 (graph C), only random blocks were included. Learning paths with fewer than two subjects or with less than two backward/forward performance scores per behavior were excluded

transitioned suddenly from one behavior to another, resulting in an improvement of their accuracy rate by 2-4 items. These findings suggest that participants in these contexts underwent an abrupt process of rule/sub-rule discovery rather than a gradual learning process involving the assignment of a category label to each stimulus. In contrast, in Experiment 2 transitions appear to be more gradual. While transitions from one behavior to another are associated with larger performance improvements, improvements within the same behavior also occur. Due to the small sample size, we cannot easily determine whether this difference with other experiments is an artifact or a result of the fixed presentation of the stimuli that could have led to the pure memorization of the correct sequence of category labels instead of the pairwise association between stimuli and category labels. Additionally, we conducted an analysis of the backward learning curve on participants in both the rule-based and similarity-based orders separately. Both groups exhibited a similar pattern, characterized by a sudden improvement in performance at the change-point.

Figure 14 offers a more detailed representation of the backward learning curves, with items belonging to distinct subcategories plotted separately. Participants in Experiments 1 and 3 exhibited a sudden improvement in performance following a transition from one behavior to another, with both the rule items (Subcategory 1) and exceptions (Subcategories 2 and 3) showing this trend. However, one exception to this pattern is observed in the backward learning curve of exceptions from participants in the 1-2 learning path, who consistently classified the exceptions with 50% accuracy. These findings suggest that transitions between behaviors may indicate sudden identification of rules/sub-rules, potentially separate from those related to the subcategories. In contrast, participants in Experiment 2 showed more gradual transitions for both the rule items (Subcategory 1) and exceptions (Subcategories 2 and 3). Once again, this observation may be attributed to the small sample size.

Discussion

Although models of category generalization are the simplest kind of models in the field of categorization, they have repeatedly shown to successfully predict participants' performance during transfer, and still serve as an essential tool for investigating cognitive processes. However, these models are not able to adapt their predictions over time, which precludes them from fitting learning data without a suitable statistical framework. Here, we address this issue by proposing a statistical method for applying transfer models to learning data.

Our first contribution includes the tailoring of the segmentation/clustering technique to allow transfer models to evolve over time. This technique arranges contiguous learning data into segments and associates a behavior to each segment. Each behavior is related to a specific set of parameters of the transfer model. Because different behaviors are generally related to different sets of parameters, the transfer model is then allowed to adapt its predictions from one segment to another. The peculiarity of the segmentation/clustering model as compared to classical segmentation models is that behaviors are shared among participants. This allows both a better estimation of the parameters of the model (segments have greater sizes) and a more accurate comparison among participants (same behaviors are available for each participant).

The advantages that arise from using this method are the following: *i)* to extend the application of transfer models to tasks in which a transfer phase is not conceived or in which feedback is always provided, *ii)* to allow a fruitful use of the learning phase by estimating the parameters of the model on the last learning segment, *iii)* to allow transfer data to be individually fit, since parameters can be estimated on the last learning segment, and (iv) to detect changes in behavior during learning without modeling a learning mechanism.

Our second contribution includes three novel experiments that investigate the impact of rule-based vs. similarity-based orders on learning speed in specific contexts. While within-category order (rule-based vs. similarity-based) was the main manipulation of interest, Experiment 1 combined a random alternation between categories with a variable presentation across blocks, Experiment 2 combined a random alternation between categories with a constant presentation across blocks, and Experiment 3 combined fully-blocked categories with a constant presentation across-blocks. Using survival analysis techniques, the rule-based order was found to be more beneficial than the similarity-based order in Experiments 2 and 3. Again, this is not surprising since a rule-plus-exceptions pattern emerges from the category structure itself. In addition, the contexts Random-Constant and Blocked-Constant were found to yield faster learning as compared to the context Random-Variable.

Our third contribution includes the application of the segmentation/clustering technique to a common transfer model (the Generalized Context Model, GCM) on our three experiments. The method found 4 learning regimes (low, medium, high, and perfect/almost perfect classification) in Experiments 1 and 2, and 3 learning regimes (low, high, and perfect/almost perfect classification) in Experiment 3. This might reflect the fact that the higher variability within Experiments 1 and 2 (both categories and across-blocks presentation were randomized in Experiment 1; only categories were randomized in Experiment 2; neither one nor the other were randomized in Experiment 3) might have strengthened the difficulty level of the task, yielding to an additional medium learning regime. The method also found 1 change-point in each experiment, meaning that participants moved from one learning regime to another during the task. By analyzing the coordinate of the change-points as a function of the experiment, we found that change-points in Experiment 1 have a higher coordinate than those in Experiments 2 and 3. This mirrors the finding that participants in Experiment 1 met the learning criterion later than those in Experiments 2 and 3. The analysis of the coordinate of the change-points as a function of presentation order (rule-based vs. similarity-based) showed that in Experiment 3 participants in the similarity-based order were characterized by change-points with a higher coordinate as compared to participants in the rule-based order. This means that the segmentation/clustering method partially detected the benefit (in terms of learning speed) of the rule-based order over the similarity-based order that has been observed in Experiments 2 and 3. We can affirm that our framework has enabled a simplified interpretation of the learning curves in terms of learning regimes, while capturing the main tendencies within the data. A more thorough analysis of the segmentation/clustering outputs using the backward learning curve revealed that participants' performance improved suddenly as they transitioned from one behavior to another. This could be interpreted as the detection of an "eureka" moment. This pattern was less prominent and more gradual in Experiment 2. However, it is uncertain whether this difference was due to the fixed presentation of stimuli that may have led to rote learning instead of abstractions, or it may be due to the small sample size. Overall, these findings suggest that participants underwent an abrupt process of rule/subrule discovery.

Perspectives and limitations

To demonstrate the effectiveness of our segmentation/clustering technique on a transfer model, we chose to use a simplified version of the GCM where attention to each dimension was fixed. While we recognize the importance of extending this technique to the full version of GCM, especially in realistic applications where some dimensions may be more relevant than others, our initial results are promising. We plan to further apply our technique to the full version of GCM as well as to other transfer models, such as Mezzadri et al.'s Ordinal General Context Model.

The segmentation/clustering method, as presented, assume s that the learning curve of every participant is affected by the same number of change-points. A natural extension of this method would be to allow each participant to have a different number of transition points. In this scenario, the selection of the number of change-points would take place for each participant before the segmentation of their learning progression. This approach may provide a more refined clustering of different types of learning.

By applying the segmentation/clustering technique to learning data we supposed that observations within a same segment are independent. If segments are sufficiently short, this hypothesis might be reasonable. However, a proper investigation of whether and when this hypothesis matches the reality is needed.

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 initial idea of applying the segmentation/clustering technique to transfer models came from P.R.-B. Experiments were designed and supervised by F.M. Data analysis and coding 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.

Declarations

Open practices statement The data for all experiments and the computer code (including the code for reproducing figures) are publicly available in Open Science Framework at https://osf.io/zv4jf/?view_only=8403629c320d4abfa0906c59443dd4ee. None of the experiments was preregistered.

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

References

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409–429.

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance, 18*(1), 50.

Busemeyer J, Diederich A (2010) Cognitive Modeling. SAGE Publications, Inc

Carvalho, P. F., & Goldstone, R. L. (2015). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? *Frontiers in Psychology, 6*, 505.

Carvalho, P. F., & Goldstone, R. L. (2022). A computational model of context-dependent encodings during category learning. *Cognitive Science, 46*(4), e13128.

Clapper, J. P., & Bower, G. H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 443–460.

Clapper, J. P., & Bower, G. H. (2002). Adaptative categorization in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 908–923.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society Series B (Methodological), 34*, 187–220.

Davies, J., Wilson, I., & Lam, W. (2005). Array cgh technologies and their applications to cancer genomes. *Chromosome Research, 13*, 237–248.

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.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature, 407*, 630–633.

Feldman, J. (2003). A catalog of boolean concepts. *Journal of Mathematical Psychology, 47*(1), 75–89.

Garner, W. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Gluck M, Bower G (1988) A configural-cue network model of classification learning. Bulletin of the Psychonomic Society 26(6).

Hayes, K. J. (1953). The backward curve: a method for the study of learning. *Psychological Review, 60*(4), 269.

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

Hupé P, Stransky N, Thiery J, Radvanyi E F andBarillot (2004) Analysis of array cgh data: from signal ratio to gain and loss of dna regions. Bioinformatics 20:3413–3422.

James A, Reynaud-Bouret P, Mezzadri G, Sargolini F, Bethus I, Muzy A (2022) Strategy inference during learning via cognitive activity-based credit assignment models. Scientific Reports 13:9408

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*, 457–481.

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

Kruschke JK (2008) Models of categorization. The Cambridge handbook of computational psychology pp 267–301

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing, 85*(8), 1501–1510.

Lieto, A. (2021). *Cognitive design for artificial minds*. Routledge.

Love, B., Medin, D., & Gureckis, T. (2004). Sustain: a network model of category learning. *Psychological Review, 111*, 309–332.

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. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.

Mezzadri, G., Laloë, T., Mathy, F., & Reynaud-Bouret, P. (2022a). Hold-out strategy for selecting learning models: application to categorization subjected to presentation orders. *Journal of Mathematical Psychology, 109*, 102691.

Mezzadri, G., Reynaud-Bouret, P., Laloë, T., & Mathy, F. (2022b). Investigating interactions between types of order in categorization. *Scientific Reports, 12*(1), 21625.

Mezzadri, G., Reynaud-Bouret, P., Laloë, T., & Mathy, F. (2022c). An order-dependent transfer model in categorization. *Journal of Mathematical Psychology, 107*, 102634.

Minda, J., & Smith, J. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 275–292.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*(1), 53.

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.

Nosofsky, R. M., Sanders, C., & Mcdaniel, M. (2018). A formal psychological model of classification applied to natural-science category learning. *Current Directions in Psychological Science, 27*, 129–135.

Olshen, A., Venkatraman, E., Lucito, R., & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics, 5*, 557–572.

Picard, F., Robin, S., Lebarbier, E., & Daudin, J. J. (2007). A segmentation/clustering model for the analysis of array cgh data. *Biometrics, 63*, 758–66.

Polk TA, Seifert CM (2002) Cognitive modeling. Boston Review

Pothos, E. M., & Wills, A. J. (2011). *Formal Approaches in Categorization*. Cambridge University Press.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive psychology, 3*(3), 382–407.

Rehder, B., & Hoffman, A. (2005). Thirty-something categorization results explained: Attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 811–829.

Rouder, J., & Ratcliff, R. (2004). Comparing categorization models. *Journal of Experimental Psychology: General, 133*, 63–82.

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

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1*, 54–87.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317–1323.

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., & Minda, J. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 3–27.

Smith, J., Tracy, J., & Murray, M. (1993). Depression and categorization. *Journal of Experimental Psychology: General, 122*, 331–346.

Smith, J. D., Minda, J. P., & Washburn, D. A. (2004). Category learning in rhesus monkeys: a study of the shepard, hovland, and jenkins (1961) tasks. *Journal of Experimental Psychology: General, 133*(3), 398.

Sun, R. (2008). *The Cambridge handbook of computational psychology*. Cambridge University Press.

Wills AJ (2013) Models of categorization, Oxford University Press, p 346–357

Zeaman, D., & House, B. J. (1963). The role of attention in retardate discrimination learning. *Handbook of mental deficiency New York: McGraw-Hill, 1*(3), 159–223.