An order-dependent transfer model in categorization

Giulia Mezzadri, Patricia Reynaud-Bouret, Thomas Laloë, Fabien Mathy

HAL Id: hal-03225670
https://hal.archives-ouvertes.fr/hal-03225670

Preprint submitted on 12 May 2021

An order-dependent transfer model in categorization

Giulia Mezzadri[a,*], Patricia Reynaud-Bouret[a], Thomas Laloë[a], Fabien Mathy[b]

[a]*Laboratoire J.A. Dieudonné UMR CNRS 7351, Université Côte d'Azur, Parc Valrose 06108 Nice Cedex 2, France.*
[b]*Laboratoire Bases, Corpus, Langage UMR CNRS 7320, Université Côte d'Azur, 24 avenue des Diables Bleus, 06357 Nice Cedex 4, France.*

Abstract

Most categorization models are insensitive to the order in which stimuli are presented. However, a vast array of studies have shown that the sequence received during learning can influence how categories are formed. In this paper, the objective was to better account for effects of serial order. We developed a model called Ordinal General Context Model (OGCM) based on the Generalized Context Model (GCM), which we modified to incorporate ordinal information. OGCM incorporates serial order as a feature along ordinary physical features, allowing it to account for the effect of sequential order as a form of distortion of the feature space. The comparison between the models showed that integrating serial order during learning in the OGCM provided the best account of classification of the stimuli in our data-sets.

Keywords: Categorization, sequencing, category transfer models, Generalized Context Model (GCM), rule-based order, similarity-based order, 5-fold cross-validation.

1. Introduction

The C major scale can convey a sense of joy or sadness, depending on the order in which the notes are played. The scale "do re mi fa sol la si" is typically

---

*corresponding author
 Email addresses:* `giulia.mezzadri@univ-cotedazur.fr` (Giulia Mezzadri), `patricia.reynaud-bouret@univ-cotedazur.fr` (Patricia Reynaud-Bouret), `thomas.laloe@univ-cotedazur.fr` (Thomas Laloë), `fabien.mathy@univ-cotedazur.fr` (Fabien Mathy)

classified as a joyful melody, while the scale "si la sol fa mi re do" is typically
classified as a somber one. A large number of studies have confirmed the intuition that the order in which information is presented can alter perception (Birnbaum et al., 2012; Carvalho & Goldstone, 2014a; Clapper, 2014; Rohrer, 2009, 2012), representation (Corcoran et al., 2011; Elio & Anderson, 1981, 1984; Mathy & Feldman, 2016; Sana et al., 2016; Zeithamova & Maddox, 2009), and learning (Bloom & Shuell, 1981; Carpenter & Mueller, 2013; Carvalho & Goldstone, 2015a; Helsdingen et al., 2011; Kornell & Bjork, 2008; Kornell et al., 2010; Lipsitt, 1961; Mathy & Feldman, 2009; Samuels, 1969; Sandhofer & Doumas, 2008; Zulkiply & Burt, 2012). More specific experimental studies have investigated how the temporal context influences the way categories are learned and encoded (Brady, 2008; Carvalho & Goldstone, 2020, 2014b; Jones & Sieck, 2003; Kang & Pashler, 2012; Li et al., 2012; Mack & Palmeri, 2015; Mcdaniel et al., 2013; Qian & Aslin, 2014; Yan et al., 2017; Zotov et al., 2011; Zulkiply et al., 2012); however, only few of them have attempted to model and account for order effects through the conception and use of computational models.

Computational models implement theories of learning by expressing the principles underlying the theories in more rigorous mathematical terms (Farrell & Lewandowsky, 2010). Moreover, the ability of models to generate quantitative predictions allows researchers to evaluate precisely how plausible the implemented theories are. Despite their advantages, only a few models have been designed to investigate how the temporal contiguity of stimuli can affect cognition. One example is the distributed Temporal Context Model (TCM) by Howard and Kahana (Howard & Kahana, 2001) that was developed to account for recency and contiguity effects in serial memory and free recall. Another example is the Memory And Contrast (MAC) model (Stewart et al., 2002) which aims at explaining the contrast effect in categorization by supposing that participants classify stimuli by estimating the relative difference between successive stimuli. A more recent example is the Sequential Attention Theory Model (SAT-M) (Carvalho & Goldstone, 2019) by Carvalho and Goldstone, in which features of a stimulus are encoded depending on how they relate to immediately preced-

ing stimuli and their category assignment. Based on the Sequential Attention Theory (SAT) (Carvalho & Goldstone, 2015b) , the SAT-M aims at explaining the diverse benefits of interleaved and blocked presentations of categories.

Despite these efforts, we think that a more general model is still lacking. We here present a new context model aiming at accounting for contiguity in categorization. Our hypothesis is that the contiguity of stimuli can distort how stimuli are perceived, which could influence both learning and transfer of category representations. Learning refers to the formation of the categories upon stimulus presentation (usually, a sample of stimuli is used repeatedly, grouped by blocks in which the stimuli are randomly permuted, until a learning criterion is reached), whereas transfer refers to the ability to categorize new stimuli. In a natural context, people continuously face new stimuli and must adapt their categories, but for simplicity reasons, the transfer phase generally follows the learning phase in a laboratory experiment. Categorization models can be divided into two classes, following this experimental logic: *i)* Learning models, which have the ability to evolve over time, which is convenient to model both learning and transfer and *ii)* Transfer models, which can only account for performance during transfer.

The model that we propose here is a transfer model. Although our model is flexible enough to account for a large variety of order effects, we here focus on how ordering the stimuli within a category can affect transfer. We here used a distinction made in the literature between rule-based vs. similarity-based presentation orders. In the similarity-based order, members of a same category are arranged in order to maximize the similarity between adjacent exemplars, while in the rule-based order, members of a same category are ordered following a "principal rule plus exceptions" structure. Research has shown that rule-based order is more beneficial than similarity-based order during both learning (Elio & Anderson, 1981; Mathy & Feldman, 2009) and transfer (Elio & Anderson, 1984; Mathy & Feldman, 2016). Although this result is particularly true because the categories which have been tested so far in these studies used stimuli involving discrete features that tend to favor the formation of rules, this dis-

3

tinction remains crucial to test models. The rationale is that a model should be more sensitive to a presentation order inspired of the mechanisms at play in the model. For instance, a rule-based model should be more sensitive to a rule-based presentation in which the non-diagnostic features are ordered randomly, and with the exceptions clearly separated from the examples obeying the main rule. On the contrary, a similarity-based model should be more sensitive to a presentation favoring similarity between contiguous examples. The idea is that in both cases, the environment would fit the internal mechanism of a model and should facilitate the extraction of the categories.

In this article, we propose a transfer model of human categorization which takes into account the order in which stimuli are presented. Different conceptions of order led to the creation of a few versions of the model. Then, we present the experiments as well as the statistical method used to compare the different versions of our model to an alternative transfer model. Finally, we show that our model best fits our data-sets and we discuss its benefits as compared to the existing models.

2. A new model of category transfer: Ordinal General Context Model (OGCM)

Our new model is an extension of the Generalized Context Model (GCM) (Medin & Schaffer, 1978; Nosofsky, 1986). Although GCM is a well-known exemplar model, we provide a rapid description of this model in order to fix the notation.

According to GCM, the probability of classifying a stimulus x as belonging to a given category A is given by the summed similarities of that stimulus to all learning exemplars of category A, divided by the summed similarities of stimulus x to all learning exemplars of all categories:

$$\mathbb{P}(A \mid x) = \frac{\sum_{a \in A \cap L} S(a, x)}{\sum_{K \in \mathcal{K}} \sum_{k \in K \cap L} S(k, x)}, \tag{1}$$

where L represents the set of learning items and $\mathcal{K}$ represents the set of all categories. The term $S(a, x)$ denotes the similarity between stimuli a and x, and

4

it is computed as an exponentially decaying function of the distance between the two stimuli:

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

where $d(a, x)$ denotes the distance between stimuli a and x; p is a positive constant; and c is a freely estimated sensitive parameter ($c \geq 0$). The distance between stimuli a and x is computed by:

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

where ω_i is the attention allocated to dimension i ($\omega_i \geq 0$ and $\sum_{i=1}^{\mathfrak{N}} \omega_i = 1$); r is a positive constant; a_i and x_i are the feature values of stimuli a and x on dimension i; and $\mathfrak{N}$ is the number of dimensions (stimuli are embedded in a $\mathfrak{N}$-dimensional psychological space).

Although GCM has proven to provide accurate predictions in the categorization domain (Nosofsky et al., 1992, 2017, 2018a,b; Rehder & Hoffman, 2005; Sanders & Nosofsky, 2020), its mathematical formulation does not integrate any component that is sensitive to different presentation orders.

2.1. Model description

The new Ordinal General Context Model (OGCM) is an extension of GCM allowing stimuli to be encoded as a function of both their physical features and their ordinal distance. With the objective of evaluating different definitions of ordinal distance, we declined OGCM into four versions: *i)* OGCM-A, which integrates the average presentation order received during learning; *ii)* OGCM-F, which integrates the most frequent presentation order received during learning; *iii)* OGCM-M, which integrates the median presentation order received during learning; and *iv)* OGCM-T, which integrates the presentation order received during transfer.

The probability of classifying a stimulus x as belonging to a given category A is defined as in Equation (1). In the same way, the similarity between two stimuli a and x is defined as in Equation (2). The single aspect that differentiates

OGCM from GCM is how the distance between two stimuli is defined. According to OGCM, the distance between stimuli a and x is determined by both physical and ordinal features as follows:

$$\tilde{d}(a,x) = \left[\sum_{i=1}^{\mathfrak{N}} \omega_i \cdot |a_i - x_i|^r + \omega_O \cdot \bar{d}_O(a,x)^r \right]^{\frac{1}{r}},$$

where the first term $\sum_{i=1}^{\mathfrak{N}} \omega_i \cdot |a_i - x_i|^r$ is the component associated with the physical features of the stimuli; $\bar{d}_O(a,x)$ is the normalized ordinal distance between stimuli a and x; ω_O is the attention allocated to ordinal features; and r is a positive constant ($\omega_i, \omega_O \geq 0$ and $\sum_{i=1}^{\mathfrak{N}} \omega_i + \omega_O = 1$). Normalizing ordinal distances allows us to compare the magnitude of the attention allocated to ordinal features to the magnitude of the attention allocated to other dimensions. The normalized ordinal distance $\bar{d}_O(a,x)$ is differently defined, depending on the version of the model. For each version of the model, we define the ordinal distance and we explain how to normalize it:

i) In OGCM-A, the ordinal distance between two learning stimuli is defined as the average temporal proximity between the two stimuli across the learning phase (see Figure 1 for an example). In order to maintain an equal comparison with the other physical dimensions, this averaged value is normalized. The normalized ordinal distance is obtained by multiplying the ordinal distance by $\frac{\max_{i=1,\ldots,\mathfrak{N}} \max_{a,x \in L \cup T} |a_i - x_i|}{|L|} \times 0.9$, where L is the set of learning stimuli and T is the set of transfer stimuli. Since the maximal temporal proximity between two learning stimuli (during learning) is equal to $|L|$, the normalized ordinal distance has an upper bound value of $\max_{i=1,\ldots,\mathfrak{N}} \max_{a,x \in L \cup T} |a_i - x_i| \times 0.9$. Multiplying this value times 0.9 allows pairs of learning stimuli to be perceived as closer as compared to pairs of stimuli in which one of them is a transfer item. Indeed, if one of the two stimuli is a transfer stimulus, then the normalized ordinal distance between them is set equal to $\max_{i=1,\ldots,\mathfrak{N}} \max_{a,x \in L \cup T} |a_i - x_i|$.

ii) In OGCM-F, the ordinal distance between two learning stimuli is defined as the most frequent temporal proximity between the two stimuli across the

learning phase (see Figure 1 for an example). This value is normalized in the same manner as in OGCM-A to maintain an equilibrium between the dimensions. When one of the stimuli is a transfer stimulus, the normalized ordinal distance is defined as in OGCM-A.

iii) In OGCM-M, the ordinal distance between two learning stimuli is defined as the median temporal proximity between the two stimuli across the learning phase (see Figure 1 for an example). Again, this value is normalized as in OGCM-A. When one of the stimuli is a transfer stimulus, the normalized ordinal distance is again defined as in OGCM-A.

iv) In OGCM-T, the ordinal distance between two stimuli (learning or transfer stimuli) is defined as the temporal proximity between the two stimuli within the considered transfer block (see Figure 1 for an example). In order to maintain an equilibrium between the dimensions, this value is multiplied by $\frac{\max\limits_{i=1,\ldots,\mathfrak{N}} \max\limits_{a,x\in L\cup T} |a_i - x_i|}{|L\cup T|}$, where L is the set of learning stimuli and T is the set of transfer stimuli. Since during the transfer phase the maximal temporal proximity between two stimuli is equal to $|L\cup T|$, the normalized ordinal distance has an upper bound value of $\max\limits_{i=1,\ldots,\mathfrak{N}} \max\limits_{a,x\in L\cup T} |a_i - x_i|$. Because we consider the temporal proximity within a transfer block, a distinction between pairs of learning stimuli and pairs of stimuli in which one of them is a transfer item is no longer needed.

3. Method

The four versions of OGCM were compared to GCM to determine which model could best account for category transfer. We now present the experiments which were used to gather data to feed the models. We then describe the statistical method used for the comparisons, and we finally present the results of this analysis.
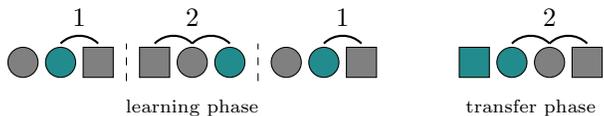
Figure 1: Example of how the ordinal distance between two stimuli is defined in OGCM, depending on its versions. We consider a simple case in which the learning phase is composed of three blocks and the transfer phase of one block. We focus our attention on computing the ordinal distance between the blue circle and the gray square. The plain curve lines and the numbers denote the temporal proximity within each block between the two selected stimuli. In OGCM-A, OGCM-F, and OGCM-M, the ordinal distance between the blue circle and the gray square is given by, respectively, the average (mean$(1, 2, 1) = 1.33$), the mode (mode$(1, 2, 1) = 1$), and the median (median$(1, 2, 1) = 1$) of their temporal proximity across the learning phase. In OGCM-T, the distance between the blue circle and the gray square is given by their ordinal position within the transfer block (i.e., 2). These values are then normalized to maintain an equilibrium between the dimensions.

3.1. Experiments

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). The procedure of this experiment is recalled in our section Experiment 1, although it is not a novel experiment. We thought that all the details needed to be reminded to the reader. This experiment was essentially designed to assess the effects of within-category orders on category transfer. Experiment 2 is a novel experiment that we carried out to use a larger number of experimental conditions than in Experiment 1.

3.1.1. Experiment 1

Experiment 1 used the famous 5-4 categories, which were first analyzed by Medin and Schaffer (Medin & Schaffer, 1978) and reanalyzed in many subsequent studies (Cohen & Nosofsky, 2003; Johansen & Kruschke, 2005; Johansen & Palmeri, 2003; Lafond et al., 2007; Lamberts, 2000; Minda & Smith, 2002; Rehder & Hoffman, 2005; Smith & Minda, 2000; Zaki et al., 2003). Because the 5-4 category structure has been evaluated several times without diminishing its current importance, it seemed like a solid starting point. In this experiment,
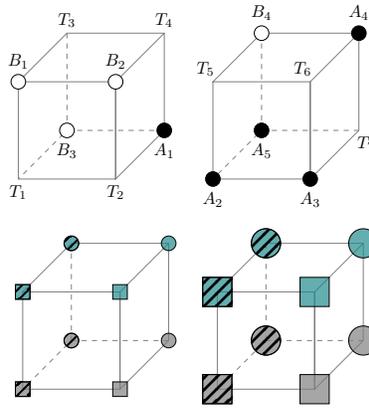
8

Figure 2: Categories and stimulus items of Experiment 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 item are represented by empty vertices. At the bottom, illustration of the items of Experiment 1.

participants were instructed to learn a 4-feature category structure based on either a rule-based presentation order or a similarity-based presentation order.

The participants ($N = 44$) were instructed to learn one category set (Figure 2, on the top), which was composed of $2^4 = 16$ items. This structure allowed to study the way in which 7 unclassified stimuli were categorized during a transfer phase, after the learning of $5+4 = 9$ stimuli. In this structure, 5 items belonged to category A and 4 items belonged to category B, for a total of $5+4 = 9$ learning items. The remaining 7 items were exclusively presented during a transfer phase. Participants were instructed to press the response key corresponding to the right category. A feedback indicated whether the chosen category was correct, except in the transfer phase.

Stimuli. Stimuli varied along four Boolean dimensions (shape, color, size, and filling pattern). The values 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 the values of the four Boolean dimensions formed $2^4 = 16$ items (Figure 2, on the bottom). Each dimension was instantiated

9

by the same physical feature for all participants. As seen in Figure 2, 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 we used to manipulate order. These training blocks were alternated with random blocks, in which stimuli were presented in a random order. The random blocks allowed to monitor the learning process while the training blocks allowed to manipulate order of the stimuli. In training blocks, order was manipulated following either a rule-based or a similarity-based order. Half of the participants were randomly assigned to the rule-based condition.

In the *rule-based* order, the stimuli were ordered following a "principal rule plus exceptions" structure, meaning that examples obeying the principal rule were presented strictly before the exceptions. The specific "principal rule plus exceptions" structure on which the rule-based order was based is the following: all gray items belong to category A except for the small hatched circle, while all blue items belong to category B except for the big plain circle (see Figure 2, page 9). Therefore, the main rule was "gray items are members of category A and 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 adjacent learning stimuli. The

10

first stimulus was randomly chosen, while subsequent stimuli were (randomly) chosen 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. The similarity-based order was thought to have the objective of reinforcing exemplar memorization (Elio & Anderson, 1981, 1984).

In training blocks, categories were strictly blocked, meaning that stimuli belonging to a same category were blocked together ($AAAABBBB$ or $BBBB$-$AAAA$). Because this repetitive order facilitated correct responses, random blocks were added to test the participants independently of our manipulation of order. In each random blocks, the nine stimuli were randomly permuted. For further details we refer the reader to (Mathy & Feldman, 2016).

3.1.2. Experiment 2

Experiment 2 was an extension of Experiment 1, but using a wider variety of types of orders. In addition to the within-category manipulations used in Experiment 1, both between-category manipulations and across-blocks manipulations were used. The introduction of these variations were thought to avoid possible interaction with the chosen types of ordering (rule-based or similarity-based). For instance, the blocked study might be beneficial for participants in the rule-based condition, but might create a disadvantage in participants in the similarity-based condition. Therefore, in comparison to Experiment 1, Experiment 2 was thought to improve the generality of our conclusions. In spite of the presence of several factors in this second experiment, the analysis of the present study only focused on the rule-based and similarity-based effects. The other factors are expected to be analyzed in a separate study.

Participants. The participants were 130 sophomore or junior students from the University Côte d'Azur (France), who received course credits in exchange for their participation.

11

Categories and stimuli. The studied categories and stimuli were the same as in Experiment 1.

Phases. As in Experiment 1, a learning phase was followed by a transfer phase. However contrary to Experiment 1, Experiment 2 made use of an unsupervised random block (i.e., participants did not get any feedback while tested). This modification was thought to better separate, at least in the participant's mind, the blocks dedicated to learning and those to testing. These unsupervised random blocks allowed us to assess learning without neither order manipulation nor feedback interfering with the measure of performance. Experiment 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. The unbalanced ratio of 2 supervised blocks followed by 1 unsupervised block aimed at increasing the influence of our manipulation, with the idea that the random block could interfere with the learning process. Participants had to correctly classify stimuli in three unsupervised blocks of 9 stimuli (not necessarily consecutive) to complete the learning phase. Once participants met the learning criterion, the transfer phase was initiated. As in Experiment 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) (Mathy & Feldman, 2016), a between-category order manipulation (interleaved vs. blocked) (Carvalho & Goldstone, 2014b), 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 ($ABABABAB$), while in the blocked order, categories were strictly blocked ($AAAABBBB$ or

12

	rule-based		similarity-based	
	interleaved	blocked	interleaved	blocked
constant	16	17	13	21
variable	14	15	15	19

Table 1: Number of participants assigned to each of the 8 conditions of Experiment 2.

$BBBBAAAA$). The introduction of random blocks during learning was thus necessary to measure the learning process, because the regularity of the patterns within blocks could hardly have participants commit classification errors. As described above, in the rule-based order, members within a category were presented following a "principal rule plus exceptions" structure, while in the similarity-based order, members within a category were presented in a way that maximized the similarity between contiguous stimuli. In the variable manipulation across blocks, the sequence of stimuli varied from one block to another (but obeying the constraints of the between- and within-category orders), while in the constant manipulation across blocks, the same sequence of stimuli (again, obeying the constraints of the between- and within-category orders) was presented in all blocks.

Procedure. The categorization task was computer-driven and was conducted online. Participants received instructions before the task began. Stimuli were presented one at a time for 3 s on the center of the computer screen. Category A was associated with the up key, while category B was associated with the down key. Participants had to classify the stimulus in one of the two categories (A and B) using these two response keys. Once the key pressed during the supervised blocks (exclusively), a feedback indicating the correctness of participants' classification appeared for 1 s at the bottom of the screen. If no key was pressed, the text 'too late' appeared for 1 s at the bottom of the screen. In order to encourage learning, the percentage of correct responses in a block was displayed for 1 s at the end of each random block.

3.2. Statistical method

Here, we first describe how the parameters of the models were estimated and we cite some numerical examples validating our estimation method. Then, we review the statistical inference method used to determine which model best accounts for category transfer. Numerical simulations were used to assess the robustness of the method.

3.2.1. Parameter Estimation

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

$$\hat{\theta} \in \underset{\theta \in \Theta}{\arg\min} \left\{ -\log \mathcal{L}_M(D\,;\theta) \right\}, \tag{3}$$

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. Numerical simulations were conducted to assess the quality of the parameter estimation as a function of the size of the data-set. The results showed that an accurate estimation of the parameters needed at least 40 blocks (for further details, see Chapter 4 of (Mezzadri, 2020)). Both experiments meet this condition, guarantying an accurate estimation of the parameters.

3.2.2. Model selection

In order to fit the studied models to the transfer phase of both Experiment 1 and Experiment 2, we used the 5-fold cross-validation method. The use of this technique was preferred to probabilistic statistical criteria, such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), because of its flexibility and robustness. The 5-fold cross-validation method is based on the holdout method, which consists in training the models on a subset of the data and testing the models on the remaining subset. Using the 5-fold cross-validation method, data are split in 5 sets and the holdout method is

14

applied 5 times. Each time, a different set is used as the testing set and the remaining $5 - 1 = 4$ sets are used as the training set. The choice of splitting the data into 5 sets is motivated by the fact that it is common advice in the literature to take either 5 or 10 sets (Arlot & Lerasle, 2016; Breiman & Spector, 1992; Hastie et al., 2009).

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:

$$E_{\text{SSD}}(M) = \sum_{x^{(t)} \in D_T} \left(\mathbb{P}_M^{\hat{\theta}} \left(A \,|\, x^{(t)} \right) - z^{(t)} \right)^2, \tag{4}$$

where M denotes the model; $\mathbb{P}_M^{\hat{\theta}} \left(A \,|\, x^{(t)} \right)$ is the prediction of the model for the stimulus $x^{(t)}$; $z^{(t)}$ 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:

$$E_{\mathcal{L}}(M) = -\log \mathcal{L}_M(D_T \,;\, \hat{\theta}), \tag{5}$$

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. Both criteria are largely used in mathematics and psychology. While in mathematics, the use of the likelihood is generally preferred, especially when the parameter estimation is performed using the MLE, in psychology, the use of the SSD is more popular than the likelihood (Carvalho & Goldstone, 2019; Nosofsky et al., 1992, 1994, 2018b; Palmeri, 1999). Both criteria were adopted to simultaneously provide a more robust evaluation and allow a continuity with previous studies.

Finally, the model characterized by the lowest averaged evaluation over the 5 application of the holdout method is the model that best fits our data-sets:

$$\hat{M} \in \underset{M}{\arg\min} \{ \underset{i=1,\ldots,5}{\text{mean}} E_*^i(M) \}, \tag{6}$$

15

where $E_*^i(M)$ is the evaluation of the model M obtained at the i-th application of the holdout method using either the SSD or the likelihood criterion. The final estimated parameter is obtained by using the MLE on the whole data-set:

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} \left\{ -\log \mathcal{L}_M(D\,;\theta) \right\}, \tag{7}$$

where M denotes the model, $\mathcal{L}_M$ its likelihood, and D the whole data-set.

3.2.3. Numerical simulations

345   Numerical simulations were conducted to assess the robustness of the method. First, the studied models were used to generate a set of artificial data, and then, the 5-fold cross-validation was applied to these artificial data-sets to determine the model that best fits them. These steps were iterated 100 times to give a statistical significance to the study. The results of the numerical simulations are
350   shown in Figure 3. The graph shows the percentage of times that the simulated data-sets were actually generated by the model that has the lowest evaluation (by means of the SSD or the likelihood criteria). Both criteria gave similar results.

If the model with the lowest evaluation was either GCM or OGCM-T, then
355   this model generated the simulated data-set with a probability of 77-82%. If the model with the lowest evaluation was either OGCM-A or OGCM-M, then this model generated the simulated data-set with a probability of 51-56%. Finally if the model with the lowest evaluation was OGCM-F, then this model generated the simulated data-set with a probability of 40%. Although OGCM-A,
360   OGCM-F, and OGCM-M were not easily distinguishable, the simulations ensure us that, when one of these three versions has the lowest evaluation, then the generative model is a model accounting for the order received during learning with a probability of 80-88%.
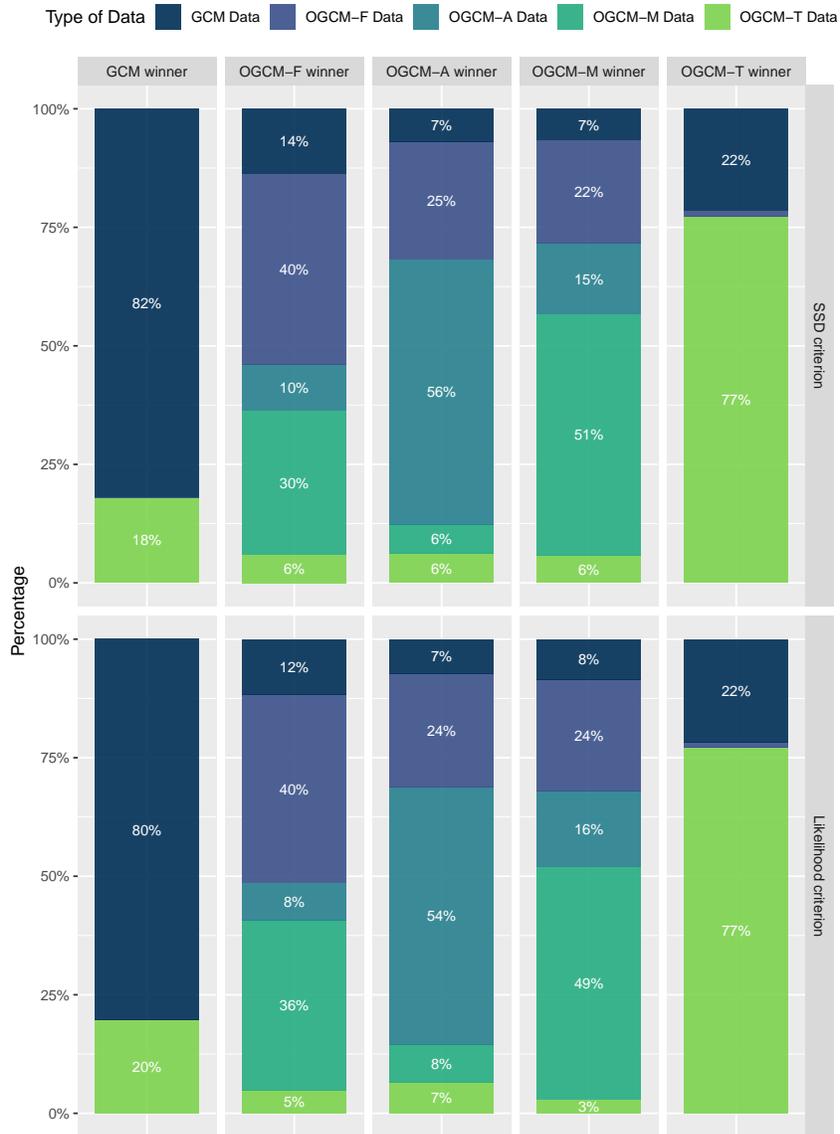
16

Figure 3: Results of the numerical simulations assessing the robustness of the method. The graph shows the percentage of times that the simulated data-sets were actually generated by the model with the lowest evaluation (using the SSD or the likelihood criteria). The procedure was iterated 100 times and a 5-fold cross-validation was adopted. Models were fit on the same sequence of stimuli used in the transfer phase of Experiment 1. The parameters used to simulate participants' responses were the result of the application of the MLE to the transfer phase of Experiment 1.

3.3. Results

3.3.1. Experiment 1

One participant (in the similarity-based order) did not meet the learning criterion and was removed from the analysis.

Fitting all participants. Table 2 (on the top) shows the model selection criteria, as well as the estimated parameters, when models were fit to the transfer phase of Experiment 1. All participants were fit together. OGCM-M and OGCM-F were the models with the lowest evaluation with the SSD criterion (SSD=91.4), and OGCM-M was the model with the lowest evaluation with the likelihood criterion ($-\log \mathcal{L} = 288.9$). The numerical simulations ensure us that the model underlying the data is a model that accounts for the order received during learning (i.e., OGCM-A, OGCM-F, or OGCM-M) with a probability of 80-88% (see Figure 3). However, the probability that the generative model is OGCM-M specifically is much lower (50%). Therefore, although we cannot determine the specific model underlying the data, we can affirm with a confidence of 80-88% that the generative model is a model that integrates serial order during learning. The estimated attention allocated to dimension 3 had the highest average value ($\hat{\omega}_3 = 0.36$ on average). This is not surprising since the third dimension corresponds to the feature that allows participants to reach the highest proportion of correct responses, which is the color. Indeed using exclusively the cue color, participants were able to correctly classify 5 of 7 learning items, which is the highest attainable value with a single cue. The last column (i.e., $\hat{\omega}_O$) is the column with the highest standard deviation ($\sigma = 0.07$). The estimated attention allocated to ordinal features was close to zero for OGCM-T ($\hat{\omega}_O = 0.06$), and far from zero for OGCM-A ($\hat{\omega}_O = 0.17$), OGCM-F ($\hat{\omega}_O = 0.20$), and OGCM-M ($\hat{\omega}_O = 0.21$). Moreover, the ordinal dimension was the second relevant dimension (after dimension 3) for OGCM-A, OGCM-F, and OGCM-M. Conversely, the ordinal dimension was the least relevant dimension for OGCM-T.

Fitting participants in the rule-based order. Table 2 (in the middle) shows the model selection criteria and the estimated parameters, when only participants

18

in the rule-based order were considered. OGCM-M and OGCM-F were the models with the lowest evaluation with the SSD criterion (SSD=43.5), while OGCM-A was the model with the lowest evaluation with the likelihood criterion $(-\log\mathcal{L} = 141.2)$. Again, the estimated attention allocated to dimension 3 had the highest average value ($\hat{\omega}_3 = 0.59$ on average). However, its average value was 1.62 times higher than the average value obtained fitting all participants. This means that the rule-based order induced participants to almost exclusively focus their attention to the dimension corresponding to the principal rule. Although the values on the other dimensions is low, we observe that the ordinal dimension was the second relevant dimension for OGCM-A, OGCM-F, and OGCM-M, while for OGCM-T was the least relevant dimension.

Fitting participants in the similarity-based order. Table 2 (on the bottom) shows the model selection criteria and the estimated parameters, when only participants in the similarity-based order were considered. OGCM-M and OGCM-F were the models with the lowest evaluation with the SSD criterion (SSD=46.0), and OGCM-M was the model with the lowest evaluation with the likelihood criterion $(-\log\mathcal{L} = 143.5)$. Again, the estimated attention allocated to dimension 3 had the highest average value ($\hat{\omega}_3 = 0.27$ on average). However, this value was 2.17 times lower than the average value obtained fitting participants in the rule-based order. This difference shows that participants in the rule-based order were highly biased toward the dimension associated with the principal rule, while participants in the similarity-based order allocated less attention to this particular dimension. Moreover, the average value of the attention allocated to dimension 4 was significantly higher as compared to the one obtained fitting participants in the rule-based order ($\hat{\omega}_4 = 0.23$ for similarity-based participants and $\hat{\omega}_4 = 0.09$ for rule-based participants). We also observe that the attention directed to the ordinal dimension had the same magnitude as the attention allocated to other dimensions, for OGCM-F ($\hat{\omega}_O$ =0.19) and OGCM-M ($\hat{\omega}_O$ =0.21). However, the attention directed to the ordinal dimension had a lower or negligible magnitude as compared to the attention allocated to other dimensions,

19

for OGCM-A ($\hat{\omega}_O$ =0.10) and OGCM-T ($\hat{\omega}_O$ =0.01). Finally, the average value

425   of the estimated sensitive parameter $\hat{c}$ was lower as compared to the one of the rule-based participants ($\hat{c} = 10.98$ in the rule-based order and $\hat{c} = 7.38$ in the similarity-based order). This means that participants in the similarity-based order classified items giving more importance to the closest items.

3.3.2. Experiment 2

430   A total of 42 participants were removed from the analysis: 6 of them (1 in the rule-based + blocked + variable condition, 2 in the rule-based + interleaved + constant, 2 in the similarity-based + blocked + variable, and 1 in the similarity-based + interleaved + variable) did not meet the learning criterion, and the remaining 36 (3 in the rule-based + interleaved + constant condition, 4 in the

435   rule-based + interleaved + variable condition, 4 in the rule-based + blocked + constant condition, 5 in the rule-based + blocked + variable condition, 1 in the similarity-based + interleaved + constant condition, 6 in the similarity-based + interleaved + variable condition, 9 in the similarity-based + blocked + constant condition, and 4 in the similarity-based + blocked + variable condition)

440   correctly classify less than 75% of the training items during the transfer phase. These participants were removed because we were interested in observing how categories were represented in participants who learned and remembered the categorization. The fact that almost 28% of participants did not remember the classification once they reached the transfer phase is probably due to running

445   the experiment online.

Regarding empty cells in the data-set, participants did not select a category in 1.4% of the analyzed data. For these trials, one of the two categories was randomly selected to facilitate modeling.

Fitting all participants. Table 3 (on the top) shows the model selection criteria,

450   as well as the estimated parameters, when models were fit to the transfer phase of Experiment 2. All participants were fit together. OGCM-F had the lowest evaluation with both criteria (SSD=175.2 and $-\log \mathcal{L} = 560.6$). Similarly to

20

the results on Experiment 1, the estimated attention allocated to dimension 3 had the highest average value ($\hat{\omega}_3 = 0.35$ on average). Moreover, the ordinal dimension was the second relevant dimension for OGCM-F ($\hat{\omega}_O$ =0.18), while for the other versions the ordinal dimension was the least relevant (for OGCM-A $\hat{\omega}_O$ =0.13, for OGCM-M $\hat{\omega}_O$ =0.11, and for OGCM-T $\hat{\omega}_O$ =0.08).

Fitting participants in the rule-based order. Table 3 (in the middle) shows the model selection criteria and the estimated parameters, when only participants in the rule-based order were considered. OGCM-A and OGCM-F had the lowest evaluation with both criteria (SSD=84.2 and $-\log\mathcal{L} = 271.0$). Again, the estimated attention allocated to dimension 3 had the highest average value ($\hat{\omega}_3 = 0.42$ on average) and its was 1.19 times higher than the average value obtained fitting all participants. This confirms that the rule-based order encouraged participants to focus their attention on the dimension corresponding to the principal rule. In OGCM-A, OGCM-F, and OGCM-M, the attention allocated to the ordinal dimension had the same magnitude as the other dimensions, with the exception of dimension 3 ($\hat{\omega}_O$ =0.15, 0.19, and 0.14, respectively). In OGCM-T, the attention allocated to the ordinal dimension had a lower magnitude as compared to all other dimensions ($\hat{\omega}_O$ =0.09).

Fitting participants in the similarity-based order. Table 3 (on the bottom) shows the model selection criteria and the estimated parameters, when only participants in the similarity-based order were considered. OGCM-A and OGCM-F were the models with the lowest evaluation with the SSD criterion (SSD=90.2), and OGCM-A was the model with the lowest evaluation with the likelihood criterion ($-\log\mathcal{L} = 287.5$). The dimension that received the highest attention were dimensions 3 and 4 ($\hat{\omega}_3 = 0.28$ and $\hat{\omega}_4 = 0.27$ on average). Similarly to Experiment 1, the average estimated attention given to dimension 3 was 1.56 times lower than its average estimated value obtained fitting participants in the rule-based order. Again, this difference shows that participants in the rule-based order were more likely to be biased toward the dimension associated with the principal rule. As in Experiment 1, the average value of the attention allocated

21

to dimension 4 was higher as compared to the one obtained fitting participants in the rule-based order ($\hat{\omega}_4 = 0.27$ for similarity-based participants and $\hat{\omega}_4 = 0.16$ for rule-based participants). We also observe that the ordinal dimension was relevant only for OGCM-A and OGCM-F ($\hat{\omega}_O$ =0.18 and 0.17, respectively). Finally inversely to Experiment 1, the average value of the estimated sensitive parameter $\hat{c}$ has the same magnitude as the one of the rule-based participants. We believe that the introduction of variations in between-category and across-blocks orders lessened the effect of the blocked study and constant manipulation across blocks on the parameter c.

4. Discussion

Previous models of category learning and transfer have rarely taken into account the order in which stimuli are actually encountered (Diamond & Kirkham, 2005; Komatsu, 1992; Murphy, 2002). Our research addressed this question, with a particular focus on within-category ordering (rule-based vs. similarity-based orders).

Our main contribution includes the development of OGCM, a model that accounts for sequencing effects by extending the usual feature space in GCM to serial order. OGCM describes how temporal proximity between stimuli "distorts" the psychological space generated by the chosen examples. We found that OGCM is the model that best fits the transfer phase of both Experiment 1 and Experiment 2. In particular, the data-set of Experiment 1 was best fit (on average) by OGCM-M, which integrates the median order received during learning, while the data-set of Experiment 2 was best fit (on average) by OGCM-F, which integrates the most frequent order received during learning. Investigating the reasons of this difference is an open interesting perspective.

In both cases, a model including ordinal features of the learning phase had the lowest model selection criterion. Moreover, the numerical simulations we conducted guarantee with a high confidence (80-88%) that the model underlying our data is a model that accounts for serial order during learning. Finally, by

analyzing the attention parameter allocated to the ordinal dimension, we found that its estimated value was significantly higher when attention was focused on ordinal information of the learning phase rather than when attention was focused on ordinal information of the transfer phase. This shows that the integration of the order in which stimuli were encountered during the learning phase can allow an ordinal model to better describe the data.

By separately fitting participants in the rule-based and similarity-based orders, we found that participants in the rule-based condition allocated greater attention to the dimension associated with the principal rule as compared to participants in the similarity-based condition (this dimension was also the most relevant for both groups). Moreover in Experiment 1, participants in the similarity-based order directed greater attention to both the 4-th and ordinal dimensions as compared to participants in the rule-based order. This shows that, in the absence of the perception of a main cue, participants in the similarity-based order relied more on ordinal information along with an additional physical feature to classify stimuli. Similarly to Experiment 1, in Experiment 2 the attention allocated to dimension 4 was higher in participants in the similarity-based order as compared to participants in the rule-based order. However, the difference in attention directed to ordinal information was no longer there. Further analyses are needed to determine whether the introduction of variations in between-category and across-blocks orders in Experiment 2 erased this difference.

Although GCM has extensively proved to provide good prediction in a large variety of contexts, it is not sensitive to sequencing effects. OGCM improves GCM by conferring the original context model the ability to adapt its predictions on the basis of the sequences of examples received by participants. This allowed OGCM to better describe the data, as well as to quantify the relevance of ordinal information on participants' classification.

As seen in the introduction, another promising modification of GCM is SAT-M. SAT-M adds to GCM four encoding weight parameters that enables the modulation of the difference between the current stimulus and the immediately preceding one. Our model is more minimalist since it extends GCM by adding

23

one more parameter. Moreover, the higher the number of parameters, the higher the number of blocks that are necessary to obtain an accurate estimate. Therefore, OGCM appears statistically more robust than SAT-M, given a data-set of a specific size.

Finally, comparing OGCM with the MAC strategy model, OGCM has a more general formulation and can be applied to a larger variety of categories. The MAC strategy also classifies items on the basis of the relative difference between the current stimulus and the immediately preceding stimulus. Therefore, this strategy can only account for the contiguity of pairs of stimuli and cannot account for the temporal proximity of the whole set of stimuli. Moreover, the MAC strategy model can exclusively be applied to mono-dimensional stimuli in which an ordinal relation between them can be defined (e.g., tones ordered from the lowest frequencies to the highest frequencies).

4.1. Limitations and Perspectives

Although OGCM was the model that best performed, the model selection criteria of the studied models were overall similar. Further analyses and different experiments are needed to pursue the investigation of the potential gain of integrating ordinal or temporal features into a model. One perspective is to compare the performance of GCM and OGCM in categorization tasks characterized by different stimuli and different categories. This comparison would allow us to determine whether the advantage of integrating the sequence of study is preserved with different materials and study conditions.

Our study was explicitly focused on transfer models and did not include learning models (Ashby et al., 1998; Erickson & Kruschke, 1998; Gluck & Bower, 1988; Kruschke, 1992; Kruschke & Johansen, 1999; Love et al., 2004; Nosofsky & Palmeri, 1998). Learning models integrate a learning mechanism that allows them to adjust their predictions over time. However, this feature does not necessarily imply their sensitivity to different types of order. Plus, sensitivity to order could exist independently of the exact learning mechanisms. A second perspective is to consider learning models and to assess their ability to predict

24

different mental representations of the learned categories as a function of the sequence being studied.

⁵⁷⁵ Because we explicitly focused on transfer models, we exclusively considered order effects during transfer. However, research has shown that different types of order can also impact the speed at which categories are learned (Elio & Anderson, 1981; Kang & Pashler, 2012; Kornell & Bjork, 2008; Mathy & Feldman, 2009). Therefore, a third perspective is to understand whether learning models ⁵⁸⁰ are capable of reproducing these findings.

Finally, our study only examined within-category orders, in particular rule-based vs. similarity-based orders. Other types of orders, such as interleaved vs. blocked study or dissimilarity-based vs. similarity-based orders, has been largely proven to impact the way we learn and represent categories as well ⁵⁸⁵ (Birnbaum et al., 2012; Carvalho & B. Albuquerque, 2012; Carvalho & Goldstone, 2014b, 2015a; Carpenter & Mueller, 2013; de Zilva & Mitchell, 2012; Kang & Pashler, 2012; Kornell & Bjork, 2008; Kornell et al., 2010; Kost et al., 2015; Kurtz & Hovland, 1956; Mathy & Feldman, 2009; Rawson et al., 2014; Rohrer, 2009, 2012; Sana et al., 2016; Taylor & Rohrer, 2010; Yan et al., 2017; ⁵⁹⁰ Wahlheim et al., 2011, 2012; Zulkiply & Burt, 2012; Zulkiply et al., 2012). A fourth perspective includes the study of the ability of the existing categorization models to predict whether different study sequences will be more or less beneficial to learners. This could significantly improve the way we adapt stimulus presentation to help learning.

⁵⁹⁵ **5. Acknowledgement**

References

Aldrich, J. (1997). R.A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, *12*, 162–176.

Arlot, S., & Lerasle, M. (2016). Choice of v for v-fold cross-validation in least-squares density estimation. *Journal of Machine Learning Research*, *17*, 1–50.

Ashby, F., Alfonso-Reese, L., Turken, A., & Waldron, E. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological review*, *105*, 442–481.

Birnbaum, M., Kornell, N., Bjork, E., & Bjork, R. (2012). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & cognition*, *41*, 392–402.

Bloom, K., & Shuell, T. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *The Journal of Educational Research*, *74*, 245–248.

Brady, F. (2008). The contextual interference effect and sport skills. *Perceptual and motor skills*, *106*, 461–72.

Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression. the x-random case. *International Statistical Review / Revue Internationale de Statistique*, *60*, 291.

Carpenter, S., & Mueller, F. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & cognition*, *41*, 671–682.

Carvalho, P., & B. Albuquerque, P. (2012). Memory encoding of stimulus features in human perceptual learning. *Journal of Cognitive Psychology*, *24*, 654–664.

Carvalho, P., & Goldstone, R. (2015a). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic bulletin & review*, *22*, 281–288.

Carvalho, P., & Goldstone, R. (2015b). 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., & Goldstone, R. (2019). A computational model of context-dependent encodings during category learning. Preprint on webpage at psyarxiv.com/8psa4/.

Carvalho, P., & Goldstone, R. (2020). The most efficient sequence of study depends on the type of test. *Applied Cognitive Psychology*, *35*, 82–97.

Carvalho, P. F., & Goldstone, R. L. (2014a). Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology*, *5*, 936.

Carvalho, P. F., & Goldstone, R. L. (2014b). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*, 481–495.

Clapper, J. (2014). The impact of training sequence and between-category similarity on unsupervised induction. *Quarterly journal of experimental psychology*, *68*, 1–55.

Cohen, A. L., & Nosofsky, R. M. (2003). An extension of the exemplar-based random-walk model to separable-dimension stimuli. *Journal of Mathematical Psychology*, *47*, 150–165.

Corcoran, K., Epstude, K., Damisch, L., & Mussweiler, T. (2011). Fast similarities: Efficiency advantages of similarity-focused comparisons. *Journal of experimental psychology. Learning, memory, and cognition*, *37*, 1280–1286.

Diamond, A., & Kirkham, N. (2005). Not quite as grown-up as we like to think parallels between cognition in childhood and adulthood. *Psychological science*, *16*, 291–297.

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.

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science*, *19*, 329–335.

Gluck, M., & Bower, G. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language - J MEM LANG*, *27*, 166–195.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* volume 1.

Helsdingen, A., Gog, T., & Van Merrienboer, J. J. G. (2011). The effects of practice schedule on learning a complex judgment task. *Learning and Instruction*, *21*, 126–136.

Howard, M., & Kahana, M. (2001). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269–299.

Johansen, M., & Kruschke, J. (2005). Category representation for classification and feature inference. *Journal of experimental psychology. Learning, memory, and cognition*, *31*, 1433–1458.

Johansen, M., & Palmeri, T. (2003). Are there representational shifts in category learning? *Cognitive psychology*, *45*, 482–553.

28

Jones, M., & Sieck, W. (2003). Learning myopia: An adaptive recency effect in category learning. *Journal of experimental psychology. Learning, memory, and cognition*, *29*, 626–640.

Kang, S., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*, 97–103.

Komatsu, L. (1992). Recent views of conceptual structure. *Psychological Bulletin*, *112*, 500–526.

Kornell, N., & Bjork, R. (2008). Learning concepts and categories: is spacing the "enemy of induction"? *Psychological science*, *19*, 585–592.

Kornell, N., Castel, A., Eich, T., & Bjork, R. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and aging*, *25*, 498–503.

Kost, A. S., Carvalho, P. F., & Goldstone, R. L. (2015). Can you repeat that? the effect of item repetition on interleaved and blocked study. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society*, (pp. 1189–1194).

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). Alcove: an exemplar-based connectionist model of category learning. *Psychological review*, *99*, 22–44.

Kurtz, K., & Hovland, C. (1956). Concept learning with different sequences of instances. *Journal of experimental psychology*, *51*, 239–243.

Lafond, D., Lacouture, Y., & Mineau, G. (2007). Complexity minimization in rule-based category learning: Revising the catalog of boolean concepts and

evidence for non-minimal rules. *Journal of Mathematical Psychology*, *51*, 57–74.

Lamberts, K. (2000). Information accumulation theory of categorization. *Psychological review*, *107*, 227–260.

Li, N., Cohen, W., & Koedinger, K. (2012). Problem order implications for learning. *International Journal of Artificial Intelligence in Education*, *23*, 71–93.

Lipsitt, L. (1961). Simultaneous and successive discrimination learning in children. *Child development*, *32*, 337–347.

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

Mack, M., & Palmeri, T. (2015). The dynamics of categorization: Unraveling rapid categorization. *Journal of experimental psychology. General*, *144*, 551–569.

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.

Mcdaniel, M., Fadler, C., & Pashler, H. (2013). Effects of spaced versus massed raining in function learning. *Journal of experimental psychology. Learning, memory, and cognition*, *39*, 1417–1432.

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*.

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.

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

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

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

Nosofsky, R. M., Sanders, C., Zhu, X., & Mcdaniel, M. (2018b). 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.

Palmeri, T. (1999). Learning categories at different hierarchical levels: A comparison of category learning models. *Psychonomic bulletin & review*, *6*, 495–503.

Qian, T., & Aslin, R. (2014). Learning bundles of stimuli renders stimulus order as a cue, not a confound. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 14400–14405.

Rawson, K., Thomas, R., & Jacoby, L. (2014). The power of examples: Illustrative examples enhance conceptual learning of declarative concepts. *Educational Psychology Review*, *27*, 483–504.

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.

Rohrer, D. (2009). The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education*, *40*, 4–17.

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review*, *24*, 355–367.

Samuels, S. (1969). Effect of simultaneous versus successive discrimination training on paired-associate learning. *Journal of Educational Psychology*, *60*, 46–48.

Sana, F., Yan, V., & Kim, J. (2016). Study sequence matters for the inductive learning of cognitive concepts. *Journal of Educational Psychology*, *109*, 84–98.

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).

Sandhofer, C., & Doumas, L. (2008). Order of presentation effects in learning color categories. *Journal of Cognition and Development - J COGN DEV*, *9*, 194–221.

32

Smith, J., & Minda, J. (2000). Thirty categorization results in search of a model. *Journal of experimental psychology. Learning, memory, and cognition*, *26*, 3–27.

Stewart, N., Brown, G., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of experimental psychology. Learning, memory, and cognition*, *28*, 3–11.

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, *24*, 837–848.

Wahlheim, C., Dunlosky, J., & Jacoby, L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & cognition*, *39*, 750–763.

Wahlheim, C., Finn, B., & Jacoby, L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: Evidence for variability neglect. *Memory & cognition*, *40*, 703–716.

Yan, V., Soderstrom, N., Seneviratna, G., Bjotk, E., & Bjork, R. (2017). How should exemplars be sequenced in inductive learning? empirical evidence versus learners' opinions. *Journal of Experimental Psychology: Applied*, *23*, 403–416.

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1160–1173.

Zeithamova, D., & Maddox, W. (2009). Learning mode and exemplar sequencing in unsupervised category learning. *Journal of experimental psychology. Learning, memory, and cognition*, *35*, 731–741.

de Zilva, D., & Mitchell, C. (2012). Effects of exposure on discrimination of similar stimuli and on memory for their unique and common features. *Quarterly journal of experimental psychology (2006)*, *65*, 1123–1138.

Zotov, V., Jones, M., & Mewhort, D. (2011). Contrast and assimilation in categorization and exemplar production. *Attention, perception & psychophysics*, *73*, 621–639.

815 Zulkiply, N., & Burt, J. (2012). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & cognition*, *41*, 16–27.

Zulkiply, N., Mclean, J., Burt, J., & Bath, D. (2012). Spacing and induction: Application to exemplars presented as auditory and visual text. *Learning and*
820 *Instruction*, *22*, 215–221.

	SSD	-log$\mathcal{L}$	$\hat{c}$	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{\omega}_3$	$\hat{\omega}_4$	$\hat{\omega}_O$
GCM	93.1	293.8	7.6	0.22	0.20	0.37	0.21	-
OGCM-A	92.0	290.7	7.4	0.17	0.14	0.36	0.15	0.17
OGCM-F	91.4	289.3	7.3	0.16	0.14	0.36	0.14	0.20
OGCM-M	91.4	288.9	7.3	0.16	0.13	0.36	0.14	0.21
OGCM-T	92.9	293.5	7.7	0.20	0.19	0.37	0.19	0.06

(a) Fitting all participants

	SSD	-log$\mathcal{L}$	$\hat{c}$	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{\omega}_3$	$\hat{\omega}_4$	$\hat{\omega}_O$
GCM	44.7	144.8	8.3	0.18	0.17	0.51	0.15	-
OGCM-A	43.6	141.2	17.4	0.06	0.05	0.76	0.04	0.09
OGCM-F	43.5	141.4	10.3	0.10	0.09	0.59	0.07	0.14
OGCM-M	43.5	141.3	10.2	0.10	0.09	0.58	0.07	0.15
OGCM-T	44.6	144.7	8.7	0.16	0.15	0.51	0.13	0.06

(b) Fitting all participants in the rule-based order

	SSD	-log$\mathcal{L}$	$\hat{c}$	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{\omega}_3$	$\hat{\omega}_4$	$\hat{\omega}_O$
GCM	46.7	145.4	7.6	0.24	0.21	0.29	0.26	-
OGCM-A	46.4	144.9	7.4	0.21	0.18	0.27	0.23	0.10
OGCM-F	46.0	143.9	7.2	0.19	0.15	0.26	0.21	0.19
OGCM-M	46.0	143.5	7.1	0.19	0.14	0.25	0.21	0.21
OGCM-T	46.7	145.4	7.6	0.23	0.21	0.29	0.26	0.01

(c) Fitting all participants in the similarity-based order

Table 2: Goodness-of-fit of the GCM and the OGCM to the transfer phase of Experiment 1, as well as the estimated parameters. Models were fit to all the participants (on the top), to participants in the rule-based order (in the middle), and to participants in the similarity-based order (on the bottom). A 5-fold cross-validation was used. Parameters were estimated using the MLE and models were evaluated using either the SSD or the likelihood.

	SSD	-log$\mathcal{L}$	$\hat{c}$	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{\omega}_3$	$\hat{\omega}_4$	$\hat{\omega}_O$
GCM	176.7	564.4	7.8	0.23	0.17	0.36	0.24	-
OGCM-A	175.6	561.6	7.8	0.19	0.13	0.35	0.20	0.13
OGCM-F	175.2	560.6	8.0	0.17	0.10	0.36	0.18	0.18
OGCM-M	175.9	562.4	7.8	0.20	0.13	0.35	0.21	0.11
OGCM-T	176.1	562.8	7.9	0.21	0.15	0.35	0.22	0.08

(a) Fitting all participants

	SSD	-log$\mathcal{L}$	$\hat{c}$	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{\omega}_3$	$\hat{\omega}_4$	$\hat{\omega}_O$
GCM	85.3	274.1	7.9	0.23	0.15	0.43	0.20	-
OGCM-A	84.2	271.0	7.8	0.18	0.09	0.44	0.14	0.15
OGCM-F	84.2	271.0	7.9	0.18	0.09	0.39	0.15	0.19
OGCM-M	84.3	271.4	7.9	0.19	0.10	0.43	0.14	0.14
OGCM-T	84.8	272.9	8.1	0.20	0.13	0.41	0.17	0.09

(b) Fitting all participants in the rule-based order

	SSD	-log$\mathcal{L}$	$\hat{c}$	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{\omega}_3$	$\hat{\omega}_4$	$\hat{\omega}_O$
GCM	90.7	289.1	8.0	0.21	0.18	0.31	0.30	-
OGCM-A	90.2	287.5	8.1	0.17	0.13	0.27	0.25	0.18
OGCM-F	90.2	289.9	8.1	0.17	0.13	0.27	0.26	0.17
OGCM-M	90.5	288.5	8.0	0.18	0.15	0.28	0.27	0.12
OGCM-T	90.6	288.6	8.0	0.20	0.16	0.29	0.27	0.07

(c) Fitting all participants in the similarity-based order

Table 3: Goodness-of-fit of the GCM and the OGCM to the transfer phase of Experiment 2, as well as the estimated parameters. Models were fit to all the participants (on the top), to participants in the rule-based order (in the middle), and to participants in the similarity-based order (on the bottom). A 5-fold cross-validation was used. Parameters were estimated using the MLE and models were evaluated using either the SSD or the likelihood.