Title
The Categorization Task is Insufficient to Distinguish between Strategies: A Case for Partial-XOR-like Tasks

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Chang, Yu-Wei
Kalish, Michael

Publication Date
2023

Peer reviewed

The Categorization Task is Insufficient to Distinguish between Strategies: A Case for Partial-XOR-like Tasks

Yu-Wei Chang (ychang01@syr.edu)
Department of Psychology, Syracuse University,
430 Huntington Hall, Syracuse, NY 13244 USA

Michael L. Kalish (mlkalish@syr.edu)
Department of Psychology, Syracuse University,
430 Huntington Hall, Syracuse, NY 13244 USA

Abstract

In categorization research, competing theories are typically compared by fitting their predictions to participant's responses on a set of test items. The theory that best matches each participant's responses is identified as the strategy the participant is most likely employing. Researchers face considerable difficulty in selecting the best-fitting model due to several factors. In this study, we show the frailty of this approach. Due to pervasive model mimicry and across the similarity- and rule-based models, typical categorization task designs fail to reliably distinguish strategies. Some design modifications that might help are counter-indicated on practical grounds (e.g., carry-over effects); other possible means of improving strategy identification are also discussed.

Keywords: categorization; strategy identification; model mimicry; Bayesian estimation

Introduction

Categorization associates with our ability to understand concepts, acquire knowledge, and make predictions. Multiple theories of how people learn and generalize categories have been proposed, including categorization based on rules (Ashby & Gott 1988, Ashby & Townsend 1986), prototype similarity (Reed 1972; Rosch 1973; Smith & Minda 1998), and exemplar similarity (Estes 1986; Medin & Schaffer 1978; Nosofsky 1986). Researchers extend our knowledge of the topic by comparing performance between groups and relating differences in performance to differences between theories. In general, the aim is to identify whether the strategy[1] people are using is the one a given theory claims they have available. However, many conclusions of these studies depend heavily on how participants' performance is interpreted. For example, we can legitimately claim that a certain patient group, for example, uses less rule-based classification relative to controls only if we can identify the strategies of the patients and controls. Therefore, it is important to evaluate the techniques we use in identifying which strategy people adopt.

Strategy Identification

One way to identify strategies is to set up ideal templates from each strategy and fit the templates to participants' data. The participant is identified as adopting the strategy in which the ideal template is the most similar to the response profile. The templates may not be exclusive on every item, so the strategies often are distinguished based on only some specific items (e.g., Conaway & Kurtz, 2015; 2017).

The template approach is straightforward. However, it falls short in accounting for the variability of profiles that one strategy can generate. A strategy should represent a way to respond rather than a particular response. A wide range of possibilities would be overlooked if only few templates are used. The situation gets worse when the features are continuous, which leads to infinite possible responses.

A better way to identify strategies is to construct each strategy as a model and conduct a selection process among the models. With this multi-model approach, a set of models were fitted to each response profile and the participant is identified as adopting the strategy from the best-fit model. Models with different parameters can show the variability of the profiles and will not be limited to few templates.

This model comparison is also conducted by researchers when proposing new theories. A new model was typically compared to the existing ones not only in the field of category learning (e.g., exemplar versus prototype, Medin & Schaffer, 1978) but other areas (e.g., prospect theory versus expected utility in decision making, Kahneman & Tversky, 1979).

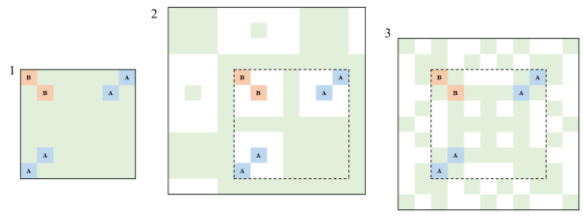

Figure 1: The category structures used in this study. The blue and orange cells are A and B exemplars respectively. The colored cells are all used as the testing items.

Partial-XOR Task

Conaway and Kurtz (2015) proposed the partial-XOR task as a way to discriminate representational strategies. The crucial feature of the task, as shown in panel 1 in Figure 1, is that

[1] We consider a 'strategy' to be identifiable with a particular distribution of responses over stimuli; a concise description of the observable phenomenon.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

items from one quadrant in an XOR structure were hidden from the participants during learning. However, in the transfer phase items from all quadrants were presented. The results showed that there were two groups of people, based on their classification of the novel items. Conaway and Kurtz interpreted a tendency to label most of these items 'B' during the transfer phase as evidence against people using similarity-based strategies in category learning.

We used the partial-XOR task as the category structure in this study not only because it plays a crucial role in the debate of similarity versus non-similarity classification, but because of its sensitivity of detecting different strategies. The transfer items, especially the items in quadrant 4 are helpful for distinguishing strategies adoption. We also included variations of the task (panel 2 and 3 in Figure 1) to examine if the multi-model-fitting approach would work better with more testing items and with the items more spread out.

Methods

To test the multi-model approach of identifying strategies, a set of rule-based models, similarity-based models, and a guessing model were collected. The parameter and model recovery analyses of these models were tested on the typical partial-XOR task and two variations of the partial-XOR task. The models were fit to the data generated not only by themselves (for parameter recovery) but also by other candidate models (for model recovery). It would serve as supporting evidence for the multi-model approach if the recovery results are stable.

The primary paradigm on researching the human category learning focuses on predicting transfer performance to new items after training. However, our interest in this paper is to consider strategy identification and not the learning performance. Since people are plausibly switching strategies back-and-forth during learning, we did not simulate the behavior in learning phase but in testing phase. Participants were assumed to memorize all the exemplars perfectly in training. In each testing trial, an item was presented, and participants were asked to classify the item into category A or B with the strategies they adopted. All testing items were presented once in random order. Along with the simulation, behavioral data on structure 1 and 2 are also reported.

Category Structures

Structure 1: the Typical Partial-XOR Task The first structure was a replication of the partial-XOR used in previous studies (e.g., Conaway & Kurtz, 2015; 2017). The stimuli varied in two features. Each feature dimension was divided into seven equal levels. Two items in quadrant 2 were B exemplars and four items in quadrant 1 and quadrant 3 were A exemplars (see panel 1 in Figure 1). A response profile was the classification of the 49 combinations of features.

Structure 2: the Extending Partial-XOR Task This structure had the same configuration of the exemplars as structure 1. Following experiment 2A in Conaway and Kurtz's study (2017), items outside the original 7x7 region were tested (see panel 2 in Figure 1). There were 12 levels in each feature dimension. The purpose is to study the extrapolation of learning on distinct testing items.

Structure 3: the Scattered Partial-XOR Task This structure had the same configuration of the exemplars as structures 1 and 2. Overall, there were 11 levels in each feature dimension. More spread items were used to obtain more information about the category strategy.

Models

Generalized Context Model (GCM) Nosofsky (1986) proposed the GCM model assuming that people represent the categories by the learned exemplars and classify items with their similarity to the exemplars. To compute the similarity, the GCM adopts a multidimensional scaling approach in that each stimulus can be represented as a point in a multi-dimensional psychological space in which the dimensions are the precepted features of the stimulus. The similarity between two items can be computed from the distance between two corresponding points in the space. The items with larger distances are less similar. Therefore, based on the GCM, when a to-be-categorized item shows up, people compute its distance to all exemplars and sum the values by categories. The categorization decision is made by choosing the category with the largest similarity to the new item, that is, the category with the smallest overall distance to the item.

$$d_{ij} = \left[\sum_{m=1}^{M} w_m \left|x_{im} - x_{jm}\right|^r\right]^{1/r} \quad (1)$$

The computation of the distance between item *i* and each exemplar *j* is shown in equation 1. *M* is the total dimensions in the psychological space. x_{im} denotes the value of item *i* on dimension *m* and x_{jm} denotes the value of exemplar *j* on the same dimension. The *w* values are attention-weight parameters with $0 \le w_m \le 1$, and $\sum w_m = 1$, reflecting the attention that people give to each dimension *m*. Items used in this study were only with two features, so *M* was equal to 2. The value *r* determines the calculation of the distance. Since the features are assumed to be psychologically separable, *r* was set to be 1, which yielded the city-block distance.

$$S_{ij} = e^{-cd_{ij}{}^p} \quad (2)$$

Equation 2 shows the formula for computing the similarity between item *i* and exemplar *j*. The value *c* is the sensitivity parameter that reflects the rate at which similarity declines with distance. The value p determines the shape of the function relating similarity to distance and was set to be 1.

$$P(J \mid i) = \frac{b_J \left[\sum_{j=1}^{n} V_{ij} S_{ij}\right]^{\gamma}}{\sum_{K=1}^{K_N} b_K \left[\sum_{k=1}^{n} V_{kK} S_{ik}\right]^{\gamma}} \quad (3)$$

Finally, the probability with which item *i* is classified into category *j* is shown in equation 3. b_J denotes the response-bias for category *j*. γ is the response-scaling parameter and was set to be 1 in this study assuming that the observer makes the decision by probability matching. *V* denotes the memory

strength of the item. In all structures used here, the memory strengths of B exemplars were set to be 2 since they were presented twice the number of trials in previous studies to maintain an equal proportion of the As and Bs in training.

The freely estimated parameters in the GCM are *c*, *w* values, and *b* values. Because the items were all 2-feature, w_2 was equal to $1 - w_1$. b_b was set to be 1 because they were all two-label categorization tasks (A or B). As a result, there were three free parameters in fitting the GCM in this study.

The GCM was included in the model set because it is a widely used and iconic similarity-based model in categorization. Besides, it is at the core of the debate whether a similarity-based model can explain the behaviors in the partial-XOR task. Previous studies showed that the GCM has some trouble fitting the XOR profiles in the task.

Similarity–dissimilarity Generalized Context Model (SD-GCM) The model was proposed by Stewart and Brown (2005). The distance computation and the similarity formula in SD-GCM are the same as the ones in GCM (Equation 1 and Equation 2). However, the SD-GCM assumes that people use not only similarity but also dissimilarity to make a category judgment. In a two-category task, if the item is sufficiently dissimilar to the members of one category, it is more likely to be classified as the other category.

$$v_{a,i} = \sum_{x_j \in A} V_{ij} S_{ij} + \sum_{x_j \in B} V_{ij} (1 - S_{ij}) \quad (4a)$$

$$v_{b,i} = \sum_{x_j \in B} V_{ij} S_{ij} + \sum_{x_j \in A} V_{ij} (1 - S_{ij}) \quad (4b)$$

With two categories A and B, the function can be denoted as Equation 4a and 4b. S is the similarity and (1-S) is the dissimilarity. V denotes the memory strength of the item and the same settings with the GCM on the parameters were applied. The SD-GCM also uses the relative rule for the category decision.

$$P(A|\, i\,) = \frac{b_a\, v_{a,i}{}^{\gamma}}{b_a\, v_{a,i}{}^{\gamma} + b_b\, v_{b,i}{}^{\gamma}} \quad (5)$$

There were three free parameters in fitting the SD-GCM in this study: *c*. w_1 and b_a. In previous studies, the SD-GCM was shown to be able to predict XOR profiles in the partial-XOR task. As a result, the SD-GCM was included in the model set not only as a similarity-based model but also as a solution within the similarity framework to address the partial-XOR question.

Cross-line Boundary Model This model assumes people use a pair of orthogonal rules to categorize items in the task. The boundaries separate the space into four regions and people assign category labels to each region. There are two parameters for the coordinate of the intersection (*x* and *y*) and four parameters for the probability of saying A in each region. If the bottom-right region has a similar probability to the up-left region and the bottom-left region has a similar probability to the up-right region, it will be an XOR profile. If only the up-left region has a different probability, it will be a proximity profile. The model was included as a representative of a rule-based model that can explain the XOR profiles (Ashby et al., 1998; Ashby & Maddox, 2005; Edmunds & Wills, 2016; Salatas & Bourne, 1974).

Hyperbolic Boundary Model This strategy model assumes that the boundary is a hyperbola. For simplicity, the degree of the hyperbola boundary is set to 45°. The formula of the hyperbolic boundary is as below:

$$(x_i - x_c)(y_i - y_c) = d \quad (6)$$

where the x_i and y_i are the positions of the item's features while the x_c and y_c are the positions of the center of the hyperbola in the corresponding dimensions. *d* is an estimated parameter that the absolute value of which indicates a scale of the distance between the lines in the hyperbola. In this study, *d* was limited to a non-positive number to consider only the top-left-to-bottom-right diagonal hyperbolas.

The hyperbolic boundary separates the space into three regions and people assign category labels to each region. There are six parameters in the model, including two for the coordinate of the center (*x* and *y*), *d*, and three for the probability of saying A in three regions. If the bottom region has a similar probability to the upper region but not the middle region, it will become an XOR profile.

The hyperbolic boundary model was included as another rule-based model. The importance of the model is that the boundaries can be represented as *one* mathematic function rather than the combination of two functions as in the previous boundary model. Thought the model is undoubtedly a decision bound theory that partitions the stimulus space into response regions, the boundaries are difficult to be described verbally, which implies feature information is integrated at some pre-decisional stage. This can be seen as a critical characteristic proposed by some researchers to be a different category learning from the cross-line boundary model (Ashby et al., 1998; Ashby & Maddox, 2005).

Guessing Model We included a guessing model as a baseline. The model has only one parameter which is the guessing rate, representing the probability of responding 'A' on each trial and is applied to every item in the task.

Bayesian Estimation for Recovery Analyses

In the study, we chose Bayesian parameter estimation instead of maximum likelihood estimation so that we would be able to explicitly incorporate priors over parameters, and so that we could estimate a mixture model to identify different groups of strategy users.

The analyses in this study were computed on R (Ihaka & Gentleman, 1996) with package rjags (Plummer et al., 2016). For each model, 100 random profiles were generated by randomly drawing parameters from a prior distribution. Each random profile was produced by applying the parameterized model to the data. The distribution of each parameter was described below. We used the same priors for random profile generation and for Bayesian model analysis.

The Priors of Parameters The positions of items in the space were normalized; therefore, they ranged from 0 to 1.

Without specific assumptions, the uniform distributions on 0 and 1 (i.e., Beta(1,1)) were used for the x, y coordinates of the intersection and the center of the hyperbola. As discussed in the model section, negative *d*s were used. When d = -1, it means that the boundary is on the edges of the space, and all testing items fall into the middle region. Thus, there is no need to include d values that are smaller than -1. As a result, the prior of d in the hyperbolic boundary model was a uniform distribution on -1 and 0 (i.e., Beta(1,1) - 1).

Without strong knowledge, the guessing parameter in the guessing model was sampled uniformly between 0 and 1. On the other hand, the priors of the region probabilities in both boundary models were set to be the beta distributions on 0.5 and 0.5 to indicate the strong belief that the probability should extremely be either 0 or 1 in each region.

For both GCM and SD-GCM, vague priors of w_1 were used (i.e., Beta(1,1)). An exponential distribution with λ = ln(2) was used as the prior of each b_a because of the desirable properties including that (1) the value is always non-negative, (2) the probability of getting an extremely big number is small, and (3) the median is 1, so the chances of preferring either category are the same. The c parameters in the GCM and SD-GCM were sampled from a Gamma distribution on 2 and 1, so extreme values were less likely to be picked, and the mode was 1.

With the parameters randomly picked, each model gave a predicted probability of being in category A on each item. A Bernoulli trial with that probability was used to generate the classification answer. Each model generated 100 profiles.

Parameter and Model Recovery Each model was fitted to the data that was generated from it, which gave out the results of recovered parameters. The parameter recovery was evaluated by examining the correlation between generating parameters and recovered parameters. Higher correlation shows a better ability of a model to find out the true parameters of the data.

Following, we conducted the model recovery. All five models were fitted to the same generated data. Three measures of fitting performance were conducted in this study. The first one was the number of counts out of 100 profiles that had best fitted by the model determined by the deviance information criterion (DIC). DIC is related to the Akaike information criteria (AIC) such that the maximized log-likelihood in the AIC is replaced by the log-likelihood evaluated at the expectation of the Bayes estimate. For each of the 100 random profiles, the model with the smallest DIC was selected as the best fit. The optimal scenario would be that each model is always selected when fitted to its own data and never selected when fitted to others.

The second measure was the average proportion of the likelihood. The likelihood was computed from ΔDIC scores as proposed by Wagenmakers and Farrell (2004). The DIC scores were scaled to be in likelihood form and turned into likelihood weights by computing the relative magnitude. i denotes different models.

$$L_{wieght,i} = \frac{L_i}{\sum L_i} \quad (7)$$

The score can range from 0 to 1 and can be interpreted as the probability that a particular model generated the observed data. The optimal scenario with the highest model distinguishability would be that each model has a score equal to 1 when fitted to its own data and 0 when fitted to others.

The third measure was the relative posterior probability for each model from a Bayesian mixture model. The model was a mixture of the five models as competing modules. Each model module applied to the data individually and had its prediction. The measure was the weight of each module in the mixture model. The weight can range from 0 to 1 and can be interpreted as the probability that a particular model module fits the data best. The optimal situation would be that each model module has a weight equal to 1 when fitted to its own data and 0 when fitted to others.

Generally, for a good model recovery result, we expect that the profiles are best fitted by the model that they are generated from and are poorly fitted by other models.

Behavioral Experiment

Participants 41 students at Syracuse University participated in the experiment. Each participant underwent three sections of classification tasks, including structure 1, structure 2, and an inside-outside structure as the filler.

Stimuli In structure 1, the stimuli were filled squares that varied in size and color. In structure 2, the stimuli were Gabor patches that varied in frequency and orientation. In the filler structure, the stimuli were blobs from Cortese and Dyre (1996) that varied in phase angle and amplitude.

Procedure Every task required participants to classify items into two categories. To minimize the confounding of memory and align with the simulation, four exemplars of each category remained labeled on-screen throughout the task. There was no training phase and participants were asked to classify presented test items regarding the given exemplars, which was similar to Yamauchi and Markman's (2000) design.

Results

Parameter Recovery

The results of correlations between generating parameters and recovered parameters are shown in Table 1. Both GCM and SD-GCM did poorly on recovering parameters c and w_1 in all three structures. Relatively, they did a better job with the response bias parameters. In the cross-line boundary model, the parameters of the region probabilities were recovered almost perfectly. The small variance may be introduced because there is no way for the model to detect the difference of boundaries smaller than the item feature interval (e.g., boundaries at 4.2 and at 4.7 perform the same when only 4 and 5 are tested). However, only the probability of the middle region was recovered in the hyperbolic boundary model. Lastly, the guessing model did well in this recovery.

Table 1: The results of parameter recovery.

Model	Parameter	Structure 1	Structure 2	Structure 3
GCM	c	.73	.62	.51
	w_1	.62	.48	.53
	b_a	.88	.93	.92
SD-GCM	c	.23	.29	.20
	w_1	.56	.48	.54
	b_a	.89	.93	.92
Cross-line boundary model	x	.79	.75	.71
	y	.68	.79	.72
	q1pa	.93	.91	.91
	q2pa	.90	.90	.89
	q3pa	.91	.94	.92
	q4pa	.93	.96	.86
Hyperbolic boundary model	x	.38	.25	.32
	y	.37	.36	.33
	uppa	.26	.28	.19
	midpa	.99	.98	.99
	botpa	.36	.45	.29
	d	.32	.46	.45
Guessing model	g	.98	.99	.98

Note. c, w, and b are sensitivity, attention weight, and response-bias respectively in GCM and SDGCM. x and y are coordinates. d is a scale of the distance between curves. The rest are the probability of response A in a certain region, including the g as a guessing rate. See text for details. The insignificant coefficients are underlined.

Model Recovery

The results of the model fitting are shown in Table 2. In each cell, the first number was how many profiles out of 100 that was best fitted (had less DIC) by the model. The second number was computed by averaging the ratio of the likelihood of the model over the sum of all models. The third number was the posterior coefficient of the model in a Bayesian mixture model, which was an indicator of the probability of the model being selected.

In a situation of a perfect model recovery, we expect to see that only the diagonal cells have numbers and others should be zeros, but as shown, it is not the case in this study. The results is similar across structures. The fitting of the GCM, the SD-GCM, the hyperbolic boundary model and the guessing model showed high missing rates that big proportions of the data generated from these models were best fitted by the others. It also shows that though the GCM and SD-GCM have different predictions on the XOR profiles, they cannot be distinguished in the partial-XOR structures.

The cross-line boundary model performed relatively better that it recovered 64, 69, and 62 out of 100 data in the three structures respectively. However, the cross-line boundary model, just as the other models, suffered from a high false alarm rate on strategy identification. The probability that data were actually generated by the cross-line boundary model when that model was identified was below 50%, which is unacceptably low. The rest of the models had worse results.

In sum, the results show problematic situations in strategy identification and model selection. Even after combining the GCM and SD-GCM together as the similarity-based models and the boundary models together as the rule-based models, the recovery was still unreliable.

Behavioral Results

The results are shown in Table 2. Across structures 1 and 2, around one-third of participants were fit best by similarity models (i.e., GCM & SDGCM), two-thirds by boundary models (i.e., cross-line & hyperbolic boundary), and none by random guessing. Since there was no direct access to what strategies they were actually using in the tasks, the recovery cannot be computed. However, the heterogeneity of model fits to individual data reflects the uncertainty we see in model recovery. We do see that the guessing model was reliably rejected as a model of people's responses, suggesting that the guessing model is too simple to compete with the other models.

Discussion

In this study, we evaluate the model-fitting approach by conducting recovery analyses with both similarity-based models and rule-based models on strategy-sensitive categorization tasks. The results show a pervasive model mimicry that models can account for the data generated by other competing models (Wagenmakers et al., 2004). Model mimicry is not directly a problem. Models should have similar prediction if they are reasonably reflective of human behavior. However, the problem appears when we use the results to prefer one model over the others. Because of the mimicry, profiles from one model were misidentified as from the others, which induced high missing rates. Besides, high false alarm rates were also observed, which indicated that the data identified as from one model may be from the others. These two errors reduce the reliability of the strategy identification, usually and especially, when the data is collected without knowing the true strategies participants use. For the same reason, it is less valid to prefer one model based on its better goodness of fit since the data might not be generated by it. We discuss the implications below.

Category Tasks Design

We used three versions of partial-XOR-like tasks as the category structures in this study. The results show that recovery was poor even when the tasks were intentionally constructed to induce different profiles.

When conducting the recovery analyses, the task designs followed the same paradigm as what is usually done in recent studies: participants first learn the exemplars of categories; then they are tested for the category generalization on new items. We also assumed the memory on the exemplars was perfect when generating data and fitting the data, which may not be true in all studies. We would expect even worse model recovery if we increased the variance by adding noise, such as errors or mistakes, when using the multi-model-fitting approach to identify strategies.

Improving the recovery results Future studies can examine if better recovery results would be obtained with some adjustments to the task designs. One possible way to improve

the recovery results and therefore have a more stable strategy identification is to increase the number of testing trials on the same item. By doing so, the classification response on each item will be a continuous scale of percentage which is more aligned to the prediction from models, compared to the dichotomous response (A or B; 1 or 0) used in this study. As a result, it would help distinguish between models. However, adding more trials inevitably increases the time-on-task, which may harm the ability on monitoring (Boksem et al., 2006). Participants may be tired and choose to or are not able to follow the strategy they used. Besides, keep asking the same item over trials may promote unsupervised or semi-supervised learning during the testing phase. This may aggravate the instability in the responses.

Another possible adjustment is to include different testing items. We attempted to do so in structure 3. However, the results show that simply spreading out testing items may not be sufficient to increase the discriminability of models. The researchers need to identify and include the crucial items that are predicted fundamentally differently between strategies. With the help of sampling-based search methods in statistics, researchers can also define the category structure to optimally discriminate models (i.e., the optimal experimental design; Myung & Pitt, 2009).

Candidate Models

We used a set of five models in this study, including two for similarity-based strategies, two for rule-based strategies, and one for the guessing strategy. The results show that the cross-line boundary model, as the one that performed the parameter recovery the best, had the smallest missing rate when fitted to its own data but still had high false alarm rate when fitted to others. This indicates that though there are connections between parameter recovery and the model comparison, parameter recovery cannot replace model recovery. Besides, the simulation results show that data generated from other models could be best fitted by the guessing model. Thus, the researchers should be cautious about whether to exclude the profiles they think as random.

One critique of the set we used may be that the SD-GCM has similar computation processes to the GCM, and this made it harder to distinguish between them. However, in our data, the model mimicry still undeniably appeared even after merging the models together according to the strategies, which says that the poor recovery results may not be induced merely because of the overlaps between models. It is worth mentioning that though we thought it is better to include more than one model in each strategy, the call of using five models was arbitrary. It is unclear whether using more or fewer models is more beneficial. Using fewer models reduces the overlaps between models and increases the discriminability; however, it falls short of exploring the range of possible strategies and induces misidentification from the forced grouping. On the other hand, including more models, as proposed by Donkin et al (2015), increases the ability to identify more potential strategies, but may induce overlap problems and overfitting problems. More studies are needed to examine how many models and which of them should be included when conducting the strategy identification.

Table 2: The results of model recovery.

Structure 1		Fitted by				
		GCM	SD-GCM	Cross-line	Hyperbolic	Guessing
Generated from	GCM	25/ **.235**/**.216**	**27**/ .227/.210	22/ .208/.188	16/ .181/.191	10/ .148/.194
	SD-GCM	21/ .199/.212	**32**/ **.277**/**.217**	28/ .234/.186	9/ .148/.190	10/ .141/.195
	Cross-line	11/ .103/.187	10/ .082/.184	**64**/ **.605**/**.260**	15/ .172/.186	0/ .039/.183
	Hyperbolic	11/ .143/.191	11/ .123/.194	19/ .207/.188	**36**/ **.315**/**.219**	23/ .212/.207
	Guessing	8/ .146/.195	16/ .181/.202	**32**/ **.241**/.185	24/ .225/.208	20/ .207/**.210**
	Behavioral Data	12/ .280/**.226**	1/ .040/.177	11/ .284/.212	**17**/ **.390**/.210	0/ .007/.175
Structure 2		Fitted by				
		GCM	SD-GCM	Cross-line	Hyperbolic	Guessing
Generated from	GCM	18/ **.202**/.212	20/ .222/**.216**	**29**/ .233/.182	16/ .171/.191	17/ .171/.199
	SD-GCM	29/ .247/.213	**23**/ **.251**/**.217**	16/ .164/.180	17/ .180/.191	15/ .158/.199
	Cross-line	6/ .056/.182	1/ .037/.181	**69**/ **.665**/**.269**	21/ .207/.187	3/ .036/.180
	Hyperbolic	8 .139/.192	8/ .110/.190	21/ .191/.182	**38**/ **.353**/**.227**	25/ .207/.208
	Guessing	21/ .194/.199	8/ .139/.200	20/ .201/.182	**30**/ **.253**/.207	21/ .212/**.212**
	Behavioral Data	9/ .222/.206	6/ .124/.197	**18**/ **.408**/**.236**	8/ .216/.183	0/ .029/.177
Structure 3		Fitted by				
		GCM	SD-GCM	Cross-line	Hyperbolic	Guessing
Generated from	GCM	27/ **.253**/.211	15/ .191/**.212**	**29**/ .238/.187	18/ .172/.191	11/ .146/.199
	SD-GCM	18/ .206/.210	**29**/ **.259**/**.216**	17/ .177/.183	23/ .198/.193	13/ .159/.198
	Cross-line	7/ .076/.184	7/ .071/.183	**62**/ **.609**/**.267**	15/ .182/.184	9/ .062/.181
	Hyperbolic	13/ .158/.193	8/ .123/.194	23/ .211/.188	25/ **.271**/**.221**	**31**/ .238/.205
	Guessing	14/ .159/.199	13/ .176/.199	**27**/ **.236**/.187	23/ .222/.207	23/ .206/**.208**

Note. In each cell, the top number is the counts of profiles that had best fitted, the lower left is the average proportion of likelihood, and the lower right is the relative posterior probability. The biggest measures in each row are bolded.

Improving the recovery results As discussed above, further studies may discover a better set of models for strategy identification. Also, there might be newly proposed models that have higher distinguishability and therefore would alleviate the model mimicry phenomena. Another possibility is the parameter spaces in the models. In this study, we mostly chose the parameters from vague priors and the recovery results were not good. However, it could happen that when the parameters are limited in a certain range, the generated data have better discriminability than what we observed. If so, the model recovery would be improved when the parameters are sampled from that range. It requires future studies to identify the special spaces of the parameters and to examine whether people's behavioral performance can be explained by the parameters within the range.

Conclusion

Through parameter and model recovery, we show that the current categorization tasks are insufficient to distinguish between the models because of the pervasive model mimicry. We suggest that it is a general question for a wider range of studies with model comparison that researchers should evaluate the distinguishability of the cognitive tasks before interpreting the results as well as pay more attention on the individual differences within the sample.

References

Ashby, F. G., Alfonso-Reese, L. A., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological review*, *105*(3), 442.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual review of psychology*, *56*(1), 149-178.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological review*, *93*(2), 154.

Boksem, M. A., Meijman, T. F., & Lorist, M. M. (2006). Mental fatigue, motivation and action monitoring. *Biological psychology*, *72*(2), 123-132.

Conaway, N., & Kurtz, K. J. (2015). A Dissociation between Categorization and Similarity to Exemplars. In *CogSci*.

Conaway, N., & Kurtz, K. J. (2017). Similar to the category, but not the exemplars: A study of generalization. *Psychonomic bulletin & review*, *24*(4), 1312-1323.

Cortese, J. M., & Dyre, B. P. (1996). Perceptual similarity of shapes generated from Fourier descriptors. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(1), 133.

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C., & Nosofsky, R. M. (2015). Identifying strategy use in category learning tasks: a case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 933.

Edmunds, C., & Wills, A. J. (2016, August). Modeling category learning using a dual-system approach: A simulation of Shepard, Hovalnd and Jenkins (1961) by COVIS. In *CogSci*.

Estes, W. K. (1986). Array models for category learning. *Cognitive psychology*, *18*(4), 500-549.

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, *5*(3), 299-314.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263-292.

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

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

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological review*, *116*(3), 499.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Plummer, M., Stukalov, A., & Denwood, M. (2016). rjags: Bayesian graphical models using MCMC. *R package version*, *4*(6).

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

Salatas, H., & Bourne, L. E. (1974). Learning conceptual rules: III. Processes contributing to rule difficulty. *Memory & Cognition*, *2*(3), 549-553.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, memory, and cognition*, *24*(6), 1411.

Stewart, N., & Brown, G. D. (2005). Similarity and dissimilarity as evidence in perceptual categorization. *Journal of Mathematical Psychology*, *49*(5), 403-409.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic bulletin & review*, *11*(1), 192-196.

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*(1), 28-50.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 776.