Psychon Bull Rev (2017) 24:617–631
DOI 10.3758/s13423-016-1123-5

Using parameter space partitioning to evaluate a model's qualitative fit

Sara Steegen[1] · Francis Tuerlinckx[1] · Wolf Vanpaemel[1]

Published online: 25 August 2016
© Psychonomic Society, Inc. 2016

Abstract Parameter space partitioning (PSP) is a versatile tool for model analysis that detects the qualitatively distinctive data patterns a model can generate, and partitions a model's parameter space into regions corresponding to these patterns. In this paper, we propose a PSP fit measure that summarizes the outcome of a PSP analysis into a single number, which can be used for model selection. In contrast to traditional model selection methods, PSP-based model selection focuses on qualitative data. We demonstrate PSP-based model selection by use of application examples in the area of category learning. A large-scale model recovery study reveals excellent recovery properties, suggesting that PSP fit is useful for model selection.

Keywords Parameter space partitioning · Model selection · Qualitative model fit · Categorization

Introduction

Parameter space partitioning (PSP; Pitt, Kim, Navarro, & Myung, 2006) is a versatile tool for model analysis, focusing on qualitative model behavior. It makes abstraction of the quantitative details in model predictions and focuses on the more general qualitative data patterns, such as ordinal relations. PSP is implemented using a Markov chain Monte Carlo (Gilks, Richardson, & Spiegelhalter, 1996) search algorithm that samples points in the parameter space and evaluates the qualitative model behavior in every selected set of parameter values. The entire parameter space gets partitioned into regions that correspond to the different qualitative data patterns the model can generate. All points belonging to the same region generate an identical qualitative data pattern. Additionally, every region is provided with an estimate of the volume it occupies in the space, reflecting the representativeness of the corresponding data pattern to the model's behavior.

Pitt et al. (2006) showed that PSP can be used for a number of purposes. First, they applied PSP to evaluate the qualitative behavior of the human category learning model ALCOVE (Kruschke, 1992) in the Shepard, Hovland, and Jenkins (1961) task, where six different category structures are learned. The qualitative pattern of interest was the ordering of the ease with which the different structures were learned. Pitt et al. (2006) evaluated ALCOVE's behavior by inspecting different aspects of its partitioned parameter space. A count of the number of generated data patterns revealed that the variety of behavioral patterns predicted by ALCOVE was quite small, and included the empirical pattern. Pitt et al. (2006) also more closely inspected the most representative data patterns (i.e., patterns which occupied the most volume in the parameter space) and found that the orderings of the different structures in these patterns were overall quite similar to the ordering in the empirical pattern. These observations suggested that ALCOVE can account for the qualitative structure of the empirical data.

In a second application, Pitt et al. (2006) applied PSP to compare the qualitative model behavior of two localist connectionist models of speech perception, TRACE (McClelland and Elman, 1986) and Merge (Norris,

Electronic supplementary material The online version of this article (doi:10.3758/s13423-016-1123-5) contains supplementary material, which is available to authorized users.

✉ Sara Steegen
sara.steegen@kuleuven.be

[1] Faculty of Psychology and Educational Sciences, KU Leuven, University of Leuven, 3000 Leuven, Belgium

McQueen, & Cutler, 2000), in two different experimental settings. The models' partitioned parameter spaces were again compared on different aspects. Pitt et al. (2006) compared the flexibility of both models by counting the number of data patterns that both models generated, and evaluated the model mimicry by studying the number of overlapping and unique data patterns (see also Hawkins, Brown, Steyvers, & Wagenmakers, 2012, Steingroever, Wetzels, & Wagenmakers, 2013 for other applications of PSP to examine model flexibility and mimicry). Further, they compared how representative the empirical pattern was to the behavior of each model by looking at the volumes of the regions in the parameter spaces that were occupied by the empirical pattern. The data patterns were also compared by means of histograms presenting the number of mismatches between the data patterns and the empirical one, and the most representative data patterns were inspected with more scrutiny, relating the volumes of their regions to the number of mismatches. Overall, these analyses suggested that TRACE and Merge show many similarities in their qualitative behavior.

These and other applications make clear that PSP is an extremely rich source that provides a great amount of useful information about a model's qualitative behavior: Does the model generate the empirical data pattern? How representative is the empirical data pattern to the model? Does the model generate many other data patterns besides the empirical one? Do the other generated data patterns deviate a lot from the empirical one? How representative are these other data patterns to the model? This richness of information, however, also has a drawback. Particularly when PSP is used for different models and one wishes to select one of the competing models as the one that performs best in capturing the empirical qualitative pattern, it is not at all straightforward how the different pieces of information provided by a PSP analysis should be meaningfully combined. Especially when the models can generate many data patterns (e.g., thousands) and the data patterns are large (e.g., rankings of more than ten stimuli), it can become a cumbersome task to find out which model provides the best qualitative account of the data.

In this paper, we propose a convenient way of summarizing the outcome of a PSP analysis into a single number useful for model selection, referred to as PSP fit. This reflects a model's ability to account for the empirical qualitative pattern. An additional goal of this paper is to evaluate the performance of PSP-based model selection using an extensive simulation study.

The remainder of the paper is organized as follows. First, we explain how PSP can be used to select among models in five steps. Next, we show four application examples where we use PSP fit to select among two competing models in the field of category learning. Finally, we evaluate the ability of PSP fit as a model selection method using a model recovery study.

PSP based model selection in five steps

To explain PSP-based model selection, we use an example of two hypothetical models, each with two parameters, denoted by θ_1 and θ_2, applied to a response time task with five conditions. We illustrate how PSP can be used as a tool for model selection in five steps.

Step 1: Define qualitative data pattern

The first step is to define the qualitative data pattern, depending on the interest of the researcher. In the current example, we assume the main interest of the researcher is capturing the ordering of the mean response times across five conditions, and thus a data pattern is defined as the rank order of mean response times across all conditions.

Step 2: Run PSP

The second step involves running PSP using the data pattern definition from step 1 (for this we use Matlab code available on http://faculty.psy.ohio-state.edu/myung/personal/psp.html). Figure 1 illustrates the partitioned parameter spaces of the two models, showing all different data patterns (in this case, rank orders) they can generate across their parameter space. Model 1's parameter space is partitioned into seven regions, corresponding to seven different rank orders of the five conditions. In each predicted pattern, the first number reflects the rank of condition 1, the second number reflects the rank of condition 2, and so forth. Rank 1 reflects the fastest and rank 5 reflects the slowest response time. For example, data pattern (5, 1, 2, 3, 4) indicates that condition 1 evoked the slowest response time, condition 2 evoked the fastest response time, and condition 3 to 5 evoked the second, third and fourth fastest response times, respectively. This data pattern occupies about one-third of the space and is the most representative of the model behavior. Being assigned a much smaller volume of 5 %, data pattern (3, 5, 2, 4, 1) is less relevant. Model 2 generates five different rank orders, the most important ones being (3, 4, 2, 5, 1), (3, 5, 1, 4, 2), and (2, 5, 3, 4, 1).

Step 3: Assess model distinguishability

The third step involves assessing the distinguishability of the models under consideration. Assessing distinguishability is crucial, since no model selection method can be reasonably expected to disentangle two models if they

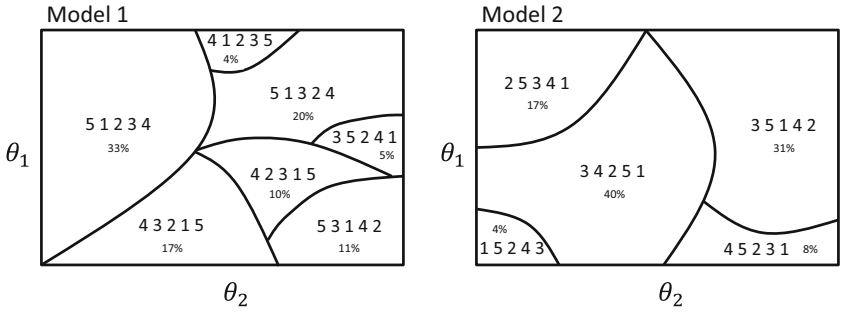

Fig. 1 The partitioned parameter spaces of two hypothetical two-parameter models in an experiment with five conditions. Model 1 generates seven distinct data patterns, whereas model 2 generates five different qualitative patterns. Every region depicts its corresponding data pattern, and the percentage of the volume that it occupies in the model's parameter space. This figure is inspired by Fig. 2 in Pitt et al. (2006)

make predictions that are very much alike (Navarro, Pitt, & Myung, 2004; Pitt, Kim, & Myung, 2003). An important benefit of using PSP to select models is that distinguishability can be assessed based on the PSP output, precluding having to perform additional simulations. A comparison of the data patterns generated by model 1 and model 2 in Fig. 1 shows that they have no data patterns in common, so the models can be distinguished.

Step 4: Collect data

Having ascertained distinguishability, it makes sense to collect empirical data. In this example, we assume that the observed response times (in ms) across the five conditions are (932, 1196, 745, 1071, 540). The qualitative representation of these data, using the data pattern definition from step 1, is (3, 5, 2, 4, 1).

Step 5: Compute PSP fit

The fifth and final step involves comparing the competing models on their ability to describe the observed data pattern. A first intuitive way of bringing the observed data pattern into contact with the predicted data patterns involves looking for the observed pattern among the predicted data patterns. The partitioned parameter spaces in Fig. 1 show that model 1 is able to generate the empirical data pattern, whereas model 2 is not. From this perspective, one could conclude that model 1 describes the observed qualitative trend better and should be preferred over model 2. The major problem with this intuitive approach is that it is highly prone to over-fitting. Over-fitting occurs when a model does not only fit variation due to the underlying process but also captures random variation. Over-fitting is usually caused by a model being too complex: An overly complex model provides a good fit to the data due to its flexibility to capture random error in the data and does not accurately reflect the underlying process. To reduce the risk of over-fitting, one should not only assess a model's ability to fit the data but also take into account the price that has to be paid to achieve this fit, in terms of the model's complexity (Myung, 2000; Pitt & Myung, 2002; Roberts & Pashler, 2000). Choosing the model that can describe the data well without being too complex is an implementation of Occam's razor.

One common solution to avoid over-fitting involves considering the average fit of a model across the entire domain of parameter values (Myung & Pitt, 1997; Pitt, Myung, & Zhang, 2002). Model complexity is taken into account in this way, in the sense that a model is penalized for being able to fit data patterns that heavily diverge from the observed data. Applied to PSP-based model selection, evaluating average model fit implies an approach that evaluates the global correspondence between the empirical data pattern and *all* the data patterns generated by the model. As explained in the introduction section, this comes down to weighing different aspects of a model's partitioned parameter space, such as how much the different data patterns deviate from the empirical one and how representative they are to the model's behavior.

The partitioned parameter spaces in Fig. 1 show that although model 1 can generate the empirical pattern, the region of model 1 corresponding to this pattern is not very representative for the model, as it only occupies a relatively small part of the parameter space. Moreover, the other data patterns generated by model 1 are rather different from the empirical one. For example, whereas the fifth and second conditions have the highest and lowest empirical ranking, respectively, this seems to be almost opposite in most data patterns generated by model 1. In contrast, the majority of the data patterns produced by model 2 show great similarities with the empirical one. Almost all pairwise relations found in the empirical data are preserved in model 2's patterns, except for some minor switches, with the most representative patterns showing the smallest number of

switches. Thus, after inspecting all generated data patterns and comparing them with the empirical pattern, it seems reasonable to conclude that globally, model 2 captures the observed qualitative trend better than model 1 does.

Whereas evaluating and comparing both models' partitioned parameter spaces in Fig. 1 is easy, in our experience, models often generate a large number of data patterns, each occupying tiny fractions of the parameter space, making it difficult to get a clear view on the models' global behavior. Especially when competing models make similar predictions, identifying the model that provides the best overall description of the empirical pattern can be challenging. Therefore, we propose to formally summarize a model's average PSP fit as follows, allowing an easy way of selecting between models based on PSP:

$$\text{PSP fit} = \frac{\sum_{i=1}^{n} d\left(DP^{\text{Emp}}, DP_i^{\text{Model}}\right) \cdot V_i}{d\left(DP^{\text{Emp}}, DP^{\text{Max}}\right)}, \quad (1)$$

where n indicates the number of data patterns generated by the model and d indicates the distance between two data patterns. DP^{Emp}, DP_i^{Model} and DP^{Max} correspond to the empirical data pattern, the ith data pattern generated by the model, and the theoretically most distant data pattern from the empirical one, respectively. Further, V_i is the proportion of the volume that data pattern i occupies in the model's parameter space.

The numerator is the summed distance (which is a positive value) between the empirical pattern and each data pattern generated by the model, including implausible ones. Each term is weighted by the proportion of the volume of the corresponding data pattern's region. It reflects the weighted average distance between a model's data patterns and the empirical data pattern. The weighted distance is then divided by its maximum value, which is equal to the distance between the empirical pattern and the theoretically most distant data pattern from the empirical one.[1] In this way, the average weighted fit is scaled to fall in the range [0, 1], with a lower value indicating a lower distance between model and data, and thus a better global fit. The minimal value of 0 is obtained if the model generates a single data pattern only, corresponding to the empirical pattern. As soon as other patterns are generated, the value of PSP fit increases. The maximal value of 1 is obtained if the model generates a single data pattern only, corresponding to the pattern most distant from the empirical one.

The exact implementation of the distance between two data patterns depends on the way patterns are defined in step 1. In the current example, as well as in the applications below, we focus on rank order patterns, so we need a measure that is suitable to assess the distance between rankings. To select from the wealth of ranking distances that are available (see e.g., Marder, 1995, for an overview), we restricted ourselves to measures that can be applied to partial rankings. Such rankings have ties, meaning there are certain items that have the same ranking, and can reasonably be expected to occur in psychological experiments (e.g., see the Shepard et al. (1961) task described earlier). Fagin et al. (2006) introduced four distance metrics that can be used to compare such partial rankings, obtained by generalizations of the well-known Kendall tau distance[2] (Kendall, 1938, 1975) and Spearman's footrule distance (Spearman, 1904; Diaconis and Graham, 1977). We will use the most easily implementable metric, which is Spearman's footrule distance generalized to partial rankings by assigning "midrank" values to the tied items or stimuli. This means that ties are replaced by the average of all tied positions in the ranking. For example, if the first and third stimulus in our example given above (see step 4) are tied, the empirical data pattern would be (2.5, 5, 2.5, 4, 1) instead of (3, 5, 2, 4, 1). Spearman's footrule distance between two (partial) rankings can then be computed as follows: $d\left(DP_1, DP_2\right) = \sum_{j=1}^{m} \left|DP_{1j} - DP_{2j}\right|$, with DP_{1j} and DP_{2j} the rank position of the jth element in the first and second data pattern, respectively.

Using this distance, the PSP fit of model 1 and model 2 equals 0.81 and 0.17, respectively (see Appendix A for computational details), indicating that model 2 provides a better overall description of the qualitative trend in the data than model 1.

By construction, PSP fit is not only sensitive to which patterns are generated but also to the proportion of the volumes of the regions these patterns occupy in the parameter space. If, for example, model 1 would generate the data patterns shown in Fig. 1, but with the volume of the empirical pattern's region equal to 88 % and the volumes of the other regions equal to 2 %, the PSP fit of model 1 would drop to 0.10 so that model 2 is preferred.

Application example

In this section, we demonstrate PSP-based model selection in the domain of category learning. We consider two category learning models that have a long history of being contrasted (see, e.g., Vanpaemel & Storms, 2010, for an overview): the generalized context model (GCM: Nosofsky, 1986) and the multiplicative prototype model

[1] The theoretically most distant data pattern from the empirical one does not necessarily need to be one of the data patterns generated by the model, so the distance between the empirical data pattern and the theoretically most distant data pattern will always be larger than zero.

[2] Not to be confused with Kendall's tau coefficient.

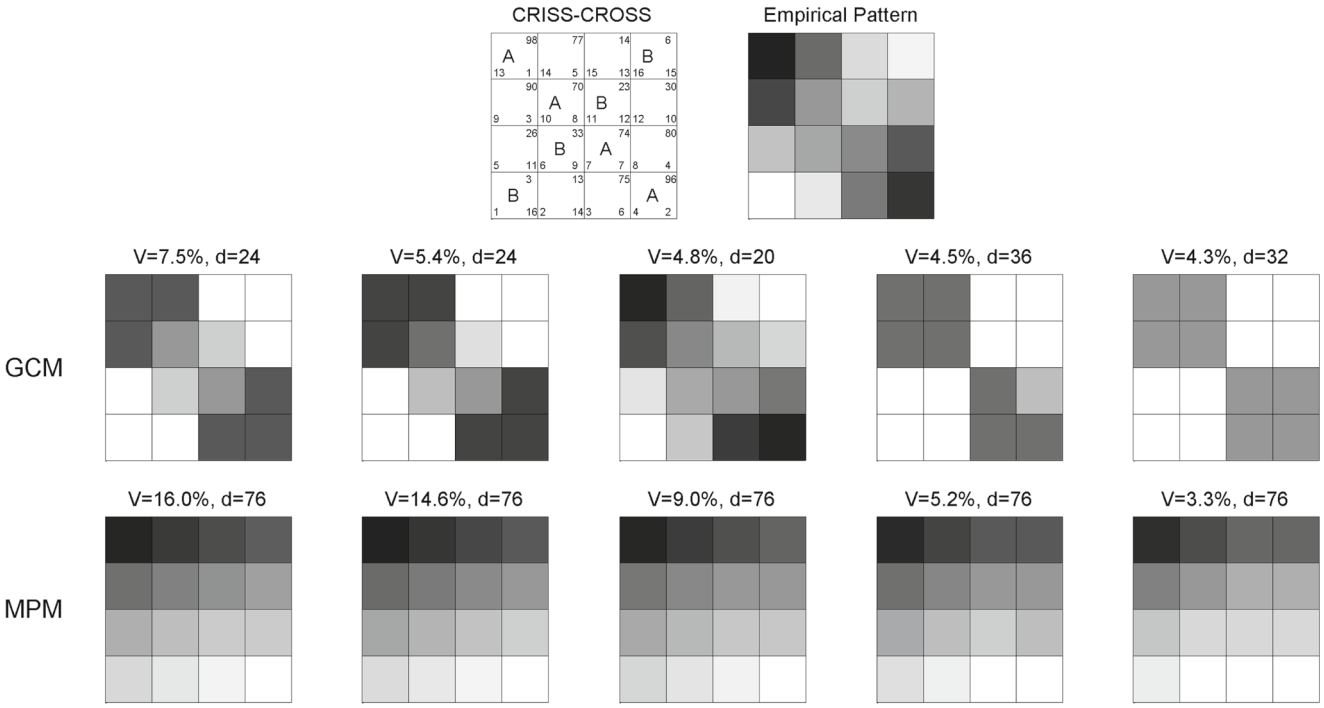

Fig. 2 The *top left* panel shows the criss-cross category structure and the data, as collected by Nosofsky (1986). *Rows and columns* reflect the four levels of the two dimensions on which the 16 stimuli vary. Cells that are marked with an A represent stimuli assigned to category A, whereas cells with a B represent stimuli assigned to category B. Cells without a letter are unassigned stimuli. Stimulus numbers are depicted in the *lower left* corner of each cell. The observed category A response percentage for each stimulus is depicted in the *upper right* corner of each cell. The corresponding rank position in the empirical data pattern is depicted in the *lower right* corner of each cell, with 1 as the highest and 16 as the lowest possible rank position. The *top right panel* shows a visualization of the empirical data pattern, where each cell in the grid represents one of the 16 stimuli to be categorized. The darker the cell, the higher the rank order position of that stimulus in the concerning pattern. The degree of darkness hence reflects the preference for category A. The *middle row* depicts a visualization of the five most representative predicted patterns of the GCM, and the *bottom row* depicts the five most representative data patterns of the MPM, using the same color coding conventions as in the visualization of the empirical data pattern. Above each predicted data pattern, the percentage of the volume (V) of the pattern and the distance (d) between the data pattern and the empirical pattern is shown

(MPM: Minda & Smith, 2011; Nosofsky, 1987; Reed, 1972). These models reflect two opposing theoretical accounts of categorization, according to which humans classify items in a certain category by comparing them to category exemplars or to a category prototype, respectively. Formal details of the models used in this application are provided in Appendix B.[3]

In a typical category learning task, participants are asked to classify previously unseen stimuli in category A or B. During the training phase, participants are trained to classify a subset of the stimuli (i.e., the assigned stimuli) in one of both categories, based on corrective feedback. During the test phase, both the assigned and the unassigned stimuli have to be categorized. In this application example, we use four category structures from Nosofsky (1986), referred to as criss-cross, interior-exterior, diagonal, and dimensional. They are based on the same set of 16 stimuli (semicircles), varying on two dimensions (size of the semicircle and angle of orientation of a radial line drawn from the center of the semicircle to the rim). The four structures are depicted schematically in the upper left panel of Figs. 2, 3, 4 and 5. Every cell in these grids schematically represents one of the 16 stimuli. Rows and columns reflect the four levels of the two dimensions on which the stimuli vary.

As a start, let us consider the criss-cross structure and apply the five steps of the PSP based model selection procedure that we have outlined above. The first step involves deciding on the data pattern of interest. In this application, we focus on the ordinal relations between the category A responses for all stimuli (e.g., stimulus 1 was classified more often in category A than stimulus 2, but less often than stimulus 3), so we define a qualitative data pattern as

[3] Our comparison between the GCM and the MPM serves as a demonstration of PSP fit as a model selection method and not as a way to draw substantial conclusions concerning the exemplar versus prototype accounts of category learning, as more recent model versions could be considered for this latter purpose (Ashby and Maddox, 1993; Vanpaemel, 2016). Further, the four data sets considered here are not intended to be representative for the vast collection of category learning data sets available, which can differ in many meaningful characteristics, including the category structures to be learned, the type of stimuli, the exact instructions, the amount of training, the type of feedback, and so on.

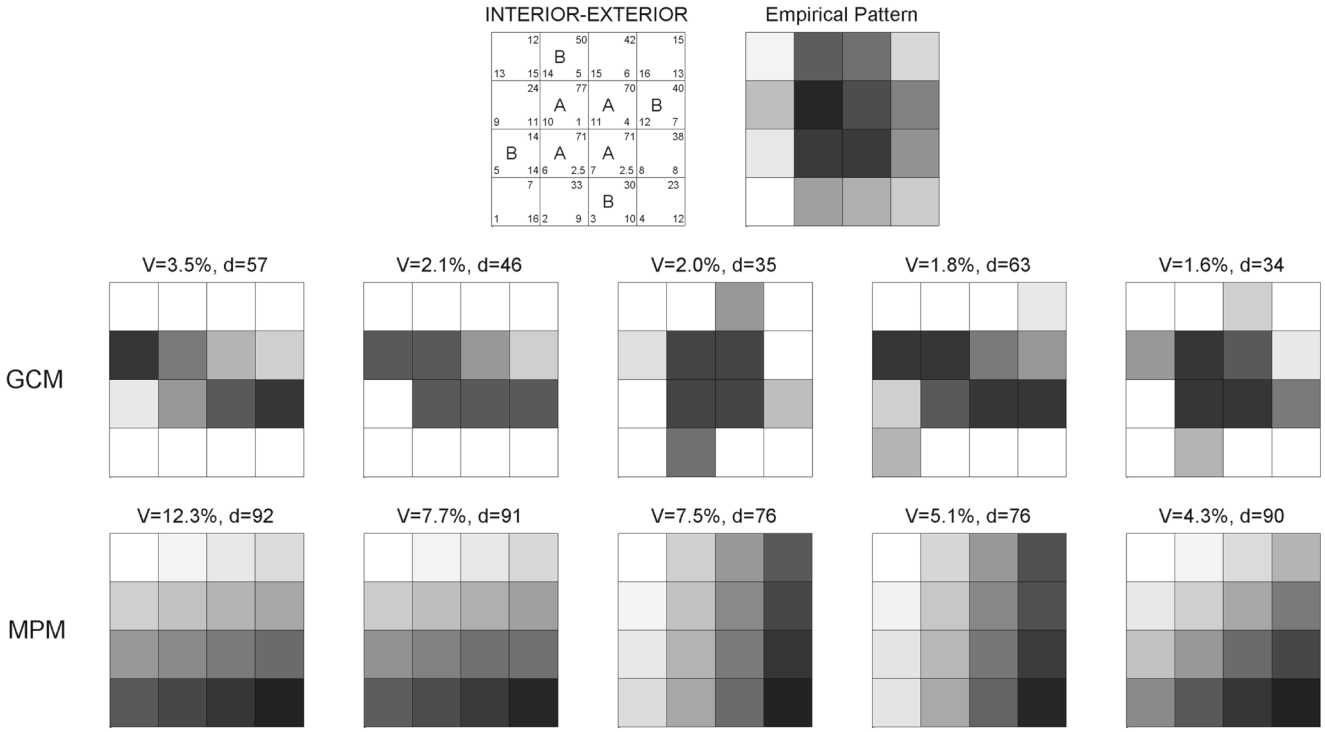

Fig. 3 Visualization of the empirical and predicted data patterns in the interior-exterior structure. See Fig. 2 for further information

the rank order of the percentage of category A responses (rounded to the nearest integer) for all stimuli. The stimulus that was assigned most often to category A gets a rank of 1, while the stimulus that was assigned most often to category B gets a rank of 16. Stimuli with equal category A response percentages get a "midrank" value.

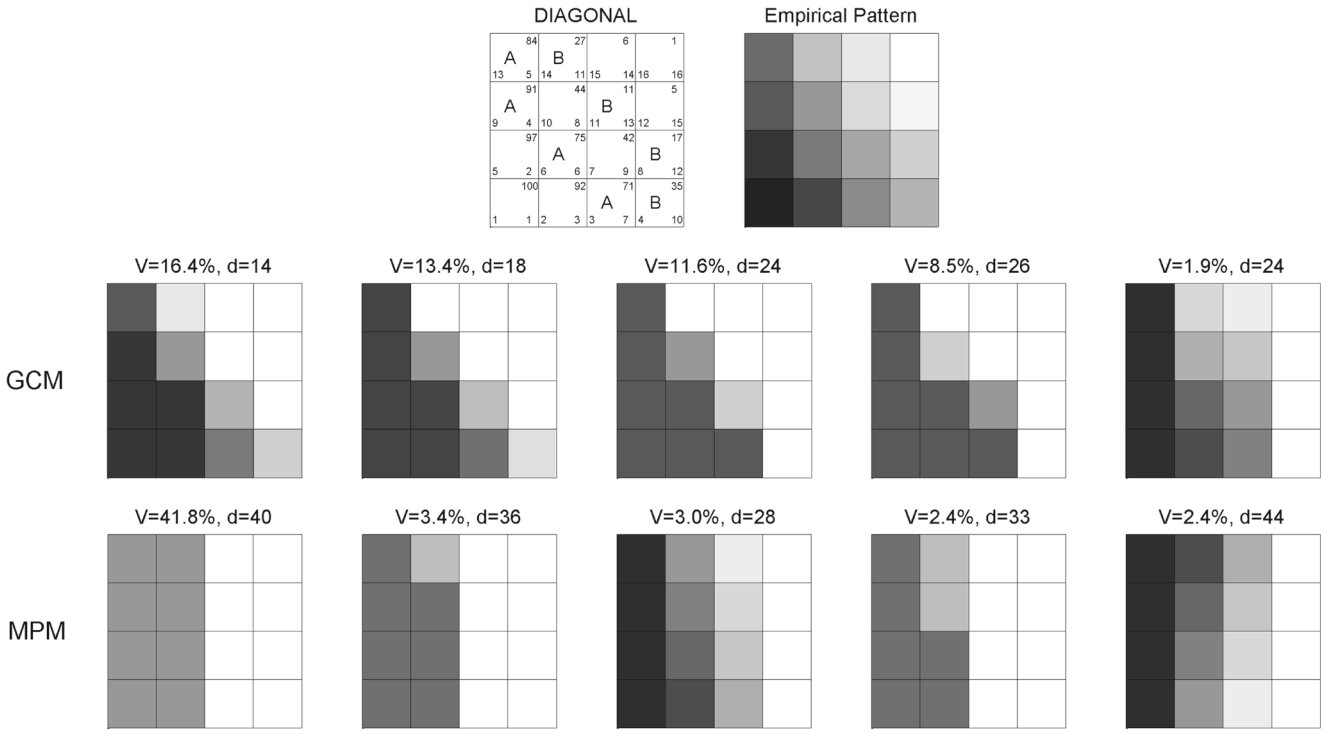

Fig. 4 Visualization of the empirical and predicted data patterns in the diagonal structure. See Fig. 2 for further information

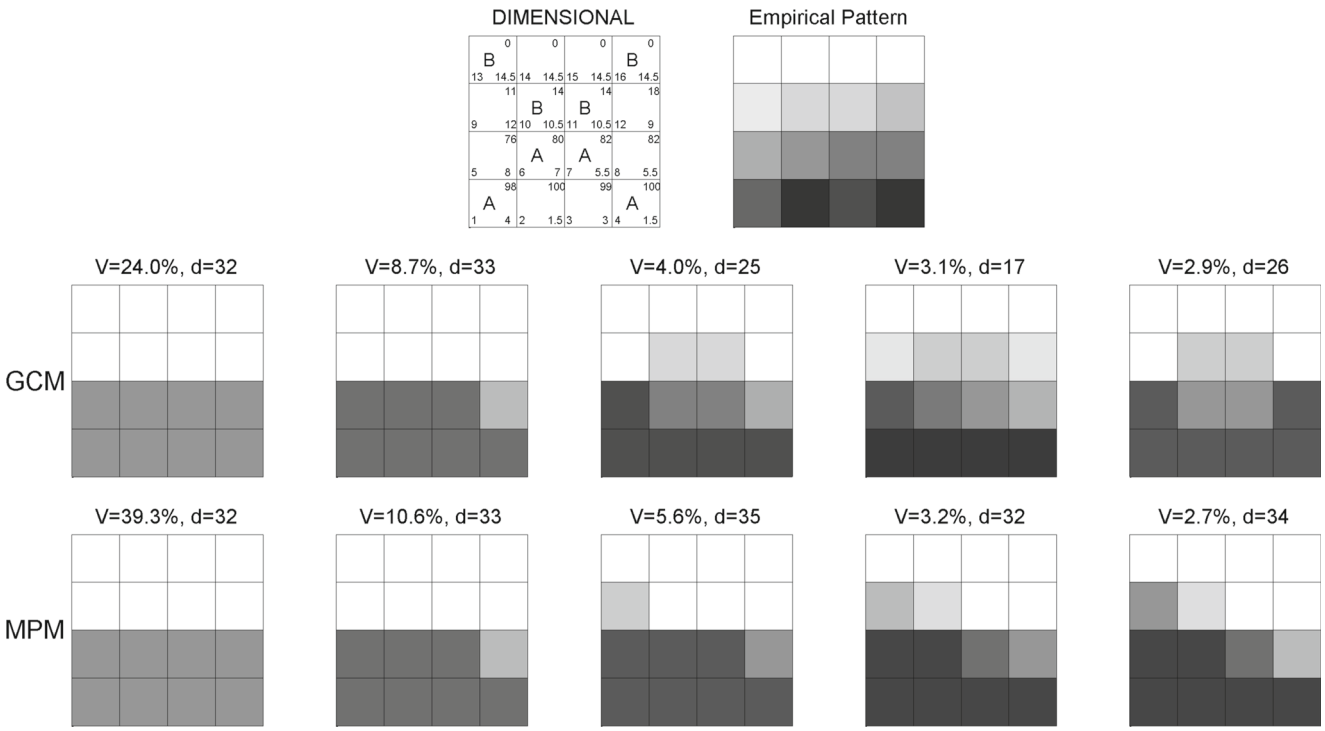

Fig. 5 Visualization of the empirical and predicted data patterns in the dimensional structure. See Fig. 2 for further information

Second, PSP is run. PSP revealed 8426 different data patterns generated by the GCM and 2142 data patterns generated by the MPM. The interested reader can find details of the PSP procedure in Appendix C.

The third step involves assessing model distinguishability. The two bottom rows in Fig. 2 show a graphical representation of the five data patterns with the largest volumes in each model's partitioned parameter space. The gradient of each cell reflects the rank position of the corresponding stimulus, with a darker gradient reflecting a higher ranking (i.e., more category A responses). This visualization provides an easy insight into the most representative qualitative model behavior and allows a quick way to roughly check model distinguishability for the largest patterns. Clearly, both models generate different types of data patterns. In the GCM patterns, stimuli that simultaneously have a low value on one and a high value on the other dimension show high rankings (dark gradient, reflecting more category A responses), whereas low or high values on both dimensions result in low rankings (light gradient, reflecting more category B responses). In contrast, rankings in the MPM patterns consistently increase with higher values on the vertical dimension and lower values on the horizontal dimension.[4] Thus, the models seem to be distinguishable.

Convinced of the models' distinguishability, it makes sense to collect data in a fourth step; or in this case, consider the data already collected by Nosofsky (1986). In this application, we will use the data obtained from Participant 1. Using the data pattern definition from step 1, the empirical qualitative data (i.e., the rankings of all stimuli) are shown in the lower right corners of the upper left panel of Fig. 2; the right panel shows a graphical representation of this qualitative data pattern. For completeness, the quantitative data—the observed percentages of category A responses, rounded to the nearest integer—are shown in the upper right corners of the cells, but these are not used in the application.

Now, the PSP fit between the empirical data pattern and the model-based data patterns can be assessed, using Eq. 1. Since we defined a data pattern as a ranking, we use Spearman's footrule distance to calculate the distances in the PSP fit, like in our introductory example. A visual comparison of the largest predicted data patterns and the empirical data pattern in Fig. 2 suggests that the predictions of the GCM provide the best qualitative fit. Calculation

[4]From the schematic representation of this category structure, these asymmetric data patterns generated by the MPM are surprising, as it seems that the prototypes of A and B have the same location (namely, in the middle of the grid). However, the model uses the psychological space of Participant 1 derived from identification data (see Fig. 5 in Nosofsky, 1986), in which the stimuli are not represented as symmetrically as in the grid representation.

of the PSP fit quantifies this observation and extends it to all 8426 and 2142 patterns produced by each model. The PSP fits of the GCM and MPM are 0.17 and 0.59, respectively (see Table 1). Since a lower value reflects a better global correspondence between the empirical data pattern and the model, these values imply that the GCM captures the observed qualitative pattern better than the MPM.

Turning to the interior-exterior structure, Fig. 3 shows the five largest data patterns generated by the GCM and MPM after running PSP (see Appendix C for more details). In total, 14,848 data patterns were generated by the GCM and 4,444 data patterns by the MPM. Visual inspection of the bottom rows of Fig. 3 reveals that both models again produce very distinguishable qualitative patterns, so it makes sense to collect data to select between both models. The five primary patterns of the GCM reflect the empirical pattern more than MPM's primary patterns do. Calculation of the PSP fits confirms that this finding generalizes to the global qualitative behavior of both models: the PSP fit equals 0.36 for the GCM and 0.65 for the MPM. Overall, GCM's performance in predicting the empirical trend in the interior-exterior structure is better than that of the MPM.

Running PSP on the diagonal structure resulted in 3,902 data patterns generated by the GCM and 6,222 data patterns generated by the MPM (see Appendix C for more details). The five most representative patterns of each model are shown in Fig. 4. Although these patterns are less dissimilar than in the previous structures, they are still distinguishable from each other. Whereas the stimuli's rankings in the GCM patterns are determined by their position on both dimensions together, the MPM mainly focuses on the stimuli's position on the horizontal dimension. Given the distinguishability, there is hope that both models can be told apart, so it is worth the effort to collect empirical data. The primary GCM patterns seem to resemble the empirical pattern more than the primary MPM patterns do. Indeed, calculation of the PSP fit leads to the selection of the GCM (PSP fit = 0.13) over the MPM (PSP fit = 0.30).

Finally, we consider the dimensional structure. Running PSP resulted in 6702 data patterns generated by the GCM and 5023 data patterns generated by the MPM (see Appendix C for more details). Figure 5 shows the models' five most representative data patterns. In this structure, the GCM and the MPM seem to closely mimic each other's behavior: the two most representative data patterns of both models, covering a rather large area in the parameter spaces, are identical. This suggests that the experimental design of the dimensional structure does not elicit sufficiently informative data to distinguish between the GCM and MPM and thus is not appropriate for disentangling these models. In this light, it seems futile to try to select between these versions of the models in this design, and one should not even bother collecting data. However, given that the data are already collected by Nosofsky (1986), we show them in Fig. 5 and we included the PSP fits in Table 1 for completeness.

Table 1 Values of the PSP fit of the GCM and the MPM in the four category structures from Nosofsky (1986)

	Model	
Structure	GCM	MPM
Criss-cross	0.17	0.59
Interior-exterior	0.36	0.65
Diagonal	0.13	0.30
Dimensional	(0.19)	(0.26)

The PSP fits in the dimensional structure are only displayed for completeness, as this structure allows very poor model distinguishability

Sensitivity analysis

Like all aspects of data analysis and modeling, PSP, and thus PSP fit, comes with several more or less arbitrary choices or researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011). First of all, there are various potential definitions of data patterns. For example, instead of considering the ranking of all stimuli, one could be interested in the pattern of A and B responses to the different stimuli (e.g., AABBA). Further, there is some freedom in choosing which stimuli to consider. Whereas we looked at all 16 stimuli that were tested, other researchers may be interested in responses to, for example, the eight unassigned stimuli only. Another researcher degree of freedom inherent to PSP concerns the constraints on the parameter space. Although some model parameters can have natural boundaries (e.g., the attention weight parameter w is restricted to [0, 1]), other parameters are unbounded (e.g., the sensitivity parameter c), so researchers have to constrain the parameter range based on what they consider appropriate. One brute but informative way of dealing with these researchers degrees of freedom is performing a sensitivity analysis, to evaluate whether and to what extent conclusions are robust across different choices.

In our sensitivity analysis, we evaluated the robustness of the above conclusions across different choices for (1) the data pattern definition, (2) the considered stimuli, and (3) the parameter range. We considered two different choices for data pattern definition: besides the ranking definition used earlier, we also considered a nominal data pattern

definition, where stimuli with more or less category A than category B responses were assigned response A or response B, respectively (e.g., AAABBAAB). To assess the distances between these data patterns for the calculation of PSP fit, we used the number of mismatches, or the Hamming distance. For example, the distance between data pattern AAAABBBB and BBBBBBBB equals 4. Concerning the considered stimuli, we examined two different choices: all 16 stimuli (as earlier) and the eight unassigned stimuli. Finally, we considered three different choices for the parameter range of c: $[0, 20]$ (as earlier), $[0, 10]$ and $[0, 5]$. In our sensitivity analysis, we combined all different choices, resulting in 12 ($2 \times 2 \times 3$) different conditions, and we calculated the PSP fits of the GCM and the MPM in each condition, for all four category learning structures, resulting in 12 tables like Table 1. As shown in Appendix D, the results are very similar. The main difference across the different conditions was that the GCM and the MPM became indistinguishable in the diagonal structure when only the unassigned stimuli were considered. Aside from that, the model selection conclusions were robust across the different choices for data patterns, considered stimuli and parameter ranges: whenever the models were distinguishable, the GCM provided better PSP fits than the MPM, in all category structures, across all 12 conditions.

Evaluation of PSP fit as a model selection tool

To evaluate the performance of PSP fit as a basis for selecting between models, we conducted an extensive model recovery study (see e.g., Navarro et al., 2004; Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). In this study, we evaluate the ability of PSP fit to recover the data-generating model. More specifically, we simulate data from both the GCM and the MPM, and assess the percentage of time in which the model selected by PSP fit corresponds to the model that generated the data.

Using the four category structures from Nosofsky (1986) as used in the application, 80,000 artificial data sets were generated. For each of the four structures, 10,000 sets of parameter values of c (with range $[0, 20]$) and w (with range $[0, 1]$) were sampled from a uniform distribution on the entire parameter space. For each set of parameter values, the probability of classifying each of the 16 stimuli in category A was computed, both according to the GCM and to the MPM. Based on these probabilities, responses were generated using the binomial distribution with the sample size that had been adopted in the concerning experiment. These responses were subsequently converted into rank order data patterns by assigning a rank position to each of the 16 stimuli, resulting in 10,000 simulated GCM data patterns and 10,000 simulated MPM data patterns. For each simulated data pattern, we calculated the PSP fit of the GCM and MPM, using Eq. 1, and we calculated a difference score: $\Delta\text{fit} = \text{PSP fit}_{\text{GCM}} - \text{PSP fit}_{\text{MPM}}$. A negative value of this difference score indicates a preference for the GCM, whereas a positive value indicates a preference for the MPM.

Next, we generated frequency distributions of the Δfit values for the patterns simulated from the GCM on the one hand and for the patterns simulated from the MPM on the other hand. Figure 6 depicts the Δfit distributions of the GCM patterns (in dark gray) and the MPM patterns (in white) in each category structure.

In the criss-cross structure, the GCM and MPM distribution lie for the greatest part at the left and at the right of $\Delta\text{fit} = 0$ (indicated by the dashed line), respectively, meaning that the majority of the simulated data patterns were correctly assigned to the model they were generated from. This observation is confirmed by the very high recovery rates, which indicate the proportion of simulated data sets that were assigned to the data-generating model (see Table 2). A small part of the MPM distribution is located at negative Δfit values, so there is a slight bias towards choosing the GCM in this structure.

A very similar picture, and similarly high recovery rates, emerge in the interior-exterior and diagonal structures. Again, the GCM and MPM distribution hardly cross $\Delta\text{fit} = 0$, so most of the data patterns were correctly assigned to the model they were generated from, based on their PSP fit.

Finally, in the dimensional structure, a large part of the GCM distribution is located at positive Δfit values (i.e., the light gray area to the right of the dashed line), and a smaller part of the MPM distribution is located at negative Δfit values (i.e., the light gray area to the left of the dashed line). This overlap is reflected in rather poor recovery rates, as shown in Table 2. These observations are consistent with our expectations based on the earlier observed indistinguishability of the models in this structure, and imply, again, that selecting between the GCM and the MPM would have been unwise in this structure, since it is impossible to know whether a GCM or MPM pattern is underlying most of the Δfit values.

Overall, for those structures where the PSP output in Figs. 2 to 4 suggested that models are distinguishable, PSP based model selection showed excellent recovery. In the one condition where a lack of distinguishability gave an early warning that model selection was elusive, model selection, unsurprisingly, turned out to be very difficult.

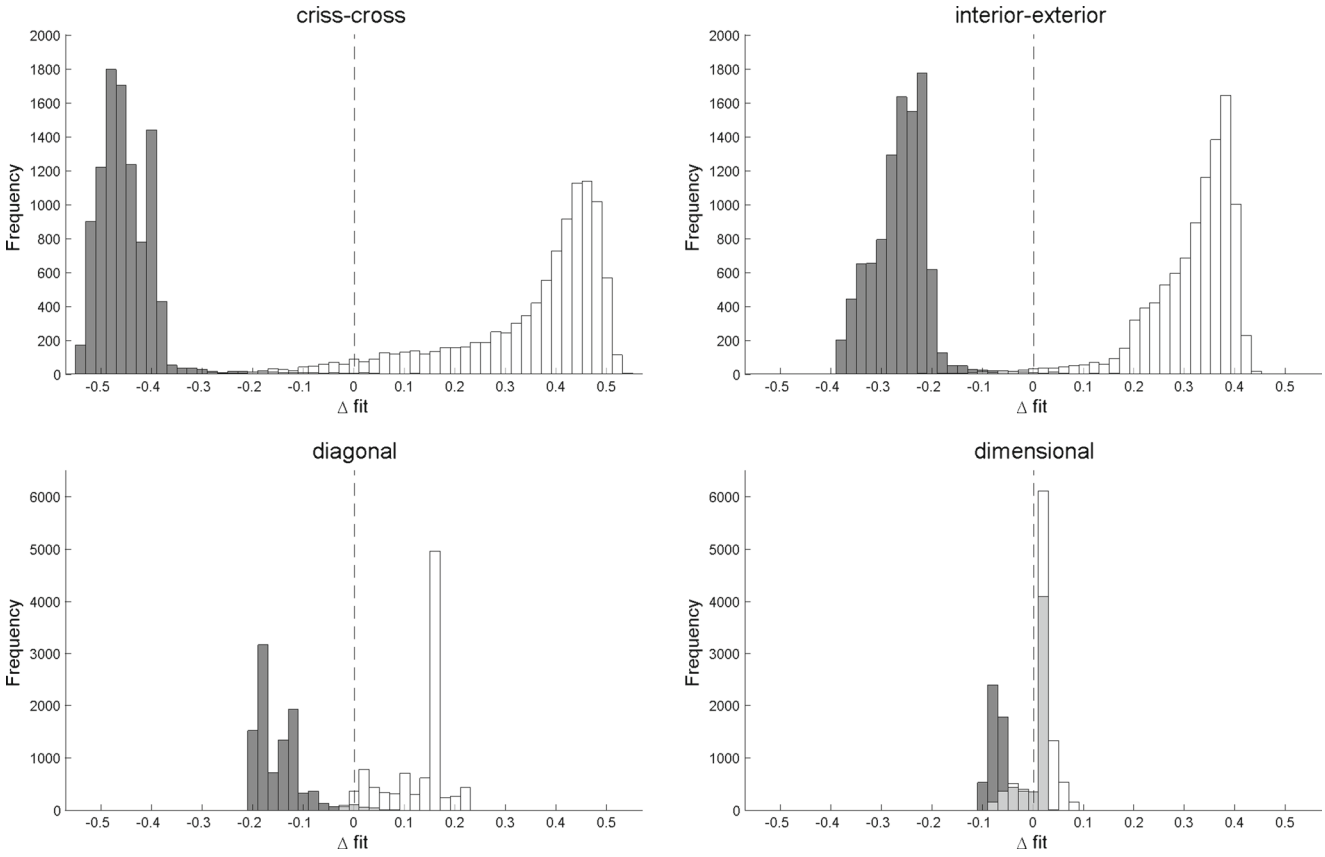

Fig. 6 Δfit distributions in the four different category structures, with $\Delta\text{fit} = \text{fit}_{\text{GCM}} - \text{fit}_{\text{MPM}}$. The *dashed line* indicates $\Delta\text{fit} = 0$. A negative value of Δfit indicates a preference for the GCM, whereas a positive value indicates a preference for the MPM. In each panel, the most left (*dark gray*) distribution reflects the Δfit distribution for the GCM patterns and the most right (*white*) distribution reflects the Δfit distribution for the MPM patterns. *Light gray areas* reflect overlapping parts of the GCM and MPM distributions

Discussion

In this paper, we proposed a way of quantitatively summarizing the output of a PSP analysis to assess a model's PSP fit, allowing for a straightforward decision rule for PSP based model selection. With an application in four category learning conditions, we demonstrated PSP based model selection between the GCM and the MPM. In three of the structures, PSP fit selected the GCM as the model with the best qualitative fit. In one structure, we refrained from selecting between the two models because an inspection of their largest data patterns suggested poor model distinguishability. The performance of PSP fit as a model selection method was evaluated with an extensive model recovery study. In those cases where a visual inspection of the PSP output suggested that models were distinguishable, PSP fit showed excellent recovery rates.

Table 2 Model recovery rates based on PSP fit for the GCM and the MPM in the four category structures from Nosofsky (1986)

	Recovery rate	
Structure	GCM	MPM
Criss-cross	.9967	.9526
Interior-exterior	.9949	.9863
Diagonal	.9762	.9794
Dimensional	.5781	.8291

PSP based model selection differs from more traditional model selection methods (e.g., RMSE, AIC, BIC, Bayes factor, maximum likelihood, MDL) in the sense that it focuses on qualitative, instead of quantitative, data and model behavior, such as the ordinal relations between conditions. Thus, considering the PSP fit to select among models is useful for those researchers whose main interest is in a model's qualitative adequacy (e.g., McFall & Townsend 1998; Pitt et al. 2003; Vanpaemel & Lee, 2012a;

Wills & Pothos, 2012). Importantly, PSP based model selection is not intended to replace traditional model selection methods that focus on the quantitative properties of the data. Rather, it could be most fruitfully seen as complementary to these methods.

PSP fit assesses a model's scaled weighted average fit. By considering the average fit, model complexity is taken into account. PSP fit penalizes a model for generating data patterns that are dissimilar from the empirical pattern, as it increases with larger distances between these patterns. However, through V_i, the weight of a data pattern's distance depends on the representativeness of the pattern to the model. Generating data patterns very distant from the empirical trend is not that damaging for a model, provided that these patterns occupy a small part of the parameter space only. Likewise, generating data patterns similar to the observed one is only favorable for a model if these patterns are substantively representative for its behavior, that is, if these patterns are assigned a high volume.

By design, the application scope of PSP fit is the same as that of PSP. Thus, as PSP, PSP fit is currently not applicable to nonstationary models (see Pitt et al. 2006). Further, it is important to keep in mind that much like other model selection measures such as AIC or BIC, PSP fit is a measure of *relative* model fit. Thus, it can be used to compare fits of competing models, but it is mute about the absolute fit of the models under consideration.

While the recovery study indicated that PSP fit can perform well, we anticipate several future developments regarding PSP fit that are possible or even desired. First, this model evaluation approach is, in spirit, similar to the marginal likelihood from Bayesian model selection (Jeffreys, 1961; Kass and Raftery, 1995; Myung & Pitt, 1997) because the weighted average fit is considered. However, in contrast to Bayesian methods, PSP has a rather weak sensitivity to prior knowledge. In psychological models, priors can be used to express theoretical assumptions (e.g., Vanpaemel & Lee, 2012b). For example, an informative prior on the attention weight of the MPM and GCM can be used to express the attention-optimization hypothesis, which posits that learners tend to distribute their attention so as to optimize their performance (Nosofsky, 1986). Accordingly, when a prior expresses theory, model evaluation should be sensitive to the prior (Vanpaemel, 2010). PSP in its current form allows researchers to express assumptions about parameter values only by adapting the range of the parameter space. Future work could modify PSP so that a full prior distribution can be placed on the parameter space. The effect of an informative prior would be that the volumes of the regions are expanded or shrunk. Accordingly, data patterns would be weighted less or more heavily when they are generated by parameter values which are considered less or more plausible. Interestingly, PSP can provide useful information for determining priors. For example, if PSP output shows that the model generates implausible data patterns, the regions in the parameter space corresponding to these data patterns can be excluded or down-weighted in the prior (Lee & Vanpaemel, 2016).

Second, in its current implementation, PSP fit is insensitive to sample size, which may yield suboptimal selection behavior in some specific cases. In particular, if there is a very large amount of data available, sampling error diminishes and the data become almost free of noise. In this case, being able to predict the empirical data pattern should prevail in model selection. Stated differently, when sample size grows to infinity, a model that can predict the empirical data pattern exactly should be selected over a model that cannot generate this pattern, regardless of its complexity. This property is akin to the criterion of consistency in quantitative model selection procedures (where it is known that BIC is consistent, but AIC is not; see e.g. Vriese, 2012). PSP fit may possibly penalize too heavily for model complexity when sample size is large. This observation has two consequences for the use of the current implementation of PSP fit. First, its use is most appropriate when sample size is small or moderate. Second, PSP fit is most suited for *explanatory*, rather than *predictive* modeling (see e.g., Shmueli, 2010). The purpose of explanatory modeling is to obtain the most accurate description of the (cognitive) process underlying the observed data. In this light, model complexity has to be penalized to overcome that a model is obtained that can fit many data patterns, without providing information about the underlying process. In predictive modeling, the goal is to predict new observations as accurately as possible, if necessary at the cost of theoretical accuracy. In this framework, often large amounts of data are available, and complex models are required to generate accurate predictions. PSP fit may not be not suited for this latter type of modeling.

A third extension is related to the observation that, as with all model selection methods, conclusions based on PSP fit depend on the data that are observed. Whereas a given experiment can produce a certain data pattern, it might well be that other experiments generate completely different data patterns, resulting in different model selection outcomes. Wills and Pothos (2012) argued to move beyond model comparisons based on single data sets, and rather compare models across a broad range of data sets. Using PSP, the empirical representativeness of data patterns based on a literature review (i.e., the frequency with which they are observed in experiments) could be compared to the representativeness of the data patterns for the competing models (see Steingroever et al. 2013). A possible extension of PSP fit involves catering to this type of model comparison across a broad range of empirical data patterns.

Finally, an important advantage of PSP based model selection is that model distinguishability, which is important for any model selection method, can be assessed without recourse to additional extensive simulation studies. The same PSP output that lies at the basis of the PSP fit can serve as a proxy for distinguishability. Thus, *before* data collection and model selection, a researcher can run PSP on the models under consideration and examine whether their data patterns are distinctive enough. When the data patterns are very similar, the researcher is warned upfront that it will be difficult to tell the competing models apart, and that it is not worth the effort to collect empirical data. In this paper, we relied on a visual representation of the most representative data patterns of the competing models to examine model distinguishability. The development of a formal way of assessing distinguishability that moves beyond visual inspection may be an interesting direction for future research.

In conclusion, we hope that the current version of PSP fit is a useful first step in using the rich information provided by PSP for selecting between models based on qualitative aspects of models and data.

Author Note We thank Mark Pitt, Andy Wills, Michael Lee and several other anonymous reviewers for their helpful comments. The research leading to the results reported in this paper was supported in part by the Research Fund of KU Leuven (GOA/15/003) and by the Interuniversity Attraction Poles programme financed by the Belgian government (IAP/P7/06).

The computational resources and services used in this work were partly provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government – department EWI.

Appendix A: Calculation of the PSP fit for the hypothetical models and data

In this section, we demonstrate how to compute the scaled average weighted PSP fits (1) of model 1 and model 2 to the empirical data pattern (3, 5, 2, 4, 1) under the Spearman's footrule distance, given by $d\,(DP_1, DP_2) = \sum_{j=1}^{m} |DP_{1j} - DP_{2j}|$, with DP_{1j} and DP_{2j} the rank position of element j in the first and second data pattern respectively. Like Fig. 1, Fig. 7 shows the models' parameter spaces partitioned into the regions that correspond to qualitatively different data patterns. In every region, the distance between the corresponding data pattern and the empirical pattern is depicted, multiplied with the proportion of the volume of the region. For example, the distance between the largest data pattern of model 1, (5, 1, 2, 3, 4), and the empirical data pattern is $|3-5| + |5-1| + |2-2| + |4-3| + |1-4| = 10$. Multiplying this distance with the volume of the corresponding region (i.e., 33 %) results in a weighted distance of 3.3. Summing all the weighted distances of model 1 yields $3.3 + 0.4 + 2.4 + 1.7 + 1.2 + 0.0 + 0.66 = 9.66$. Under Spearman's footrule distance, the maximum distance between two data patterns with an odd number of elements is $\frac{(m+1)(m-1)}{2}$, where m is the number of elements, so in this example $d(DP^{\text{Emp}}, DP^{\text{Max}}) = 12$. Dividing the weighted distance by this maximal distance gives a PSP fit of 0.81 for model 1. A similar calculation results in a PSP fit of 0.17 for model 2.

Appendix B: Formal details of the GCM and MPM

According to the generalized context model (GCM; Nosofsky, 1986), classification decisions are based on the similarity of the stimulus to the relevant categories. As categories are assumed to be represented by individual exemplars, every exemplar in that category is considered in determining the similarity to a category. When there are two categories, A and B, the probability that

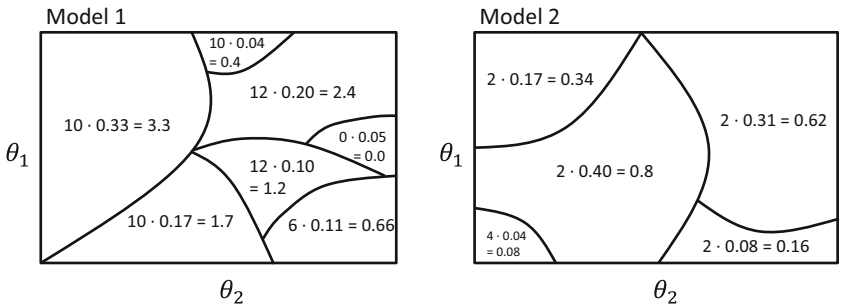

Fig. 7 The weighted distances between the fictional empirical data pattern (3, 5, 2, 4, 1) and the data patterns of hypothetical models, model 1 and model 2, as depicted in Fig. 1

stimulus i is classified into category A is formalized as:

$$P(A \mid i) = \frac{\sum_{j \in A} sim(i, j)}{\sum_{j \in A} sim(i, j) + \sum_{j \in B} sim(i, j)}, \quad (2)$$

where $sim(i, j)$ indicates the similarity between stimulus i and exemplar j, which is given by

$$sim(i, j) = \exp\left(-c\left[\sum_{m=1}^{M} w_m \mid x_{im} - x_{jm} \mid^r\right]^{\frac{1}{r}}\right). \quad (3)$$

In this equation, c is a sensitivity parameter, M is the number of stimulus dimensions, w_m is an attention-weight parameter, x_{im} and x_{jm} denote the coordinates of stimulus i and exemplar j on dimension m, respectively, and r is the metric. Most often, r is not considered a free parameter but is assumed to depend on the type of dimensions that compose the stimuli. Generally, the city block metric ($r = 1$) is used when stimuli vary on separable dimensions, and the Euclidean metric ($r = 2$) is used for stimuli varying on integral dimensions. In the applications in this paper, the Euclidean metric was used.

The multiplicative prototype model (MPM; Minda & Smith, 2011; Nosofsky, 1987; Reed, 1972) differs from the GCM in one crucial assumption: A category is assumed to be represented by one abstract summary item, which means that, in determining the similarity of a stimulus to a category, only this item is considered. This summary item, called the prototype, is defined as the average of all individual exemplars. The coordinates of the category A prototype are the averaged coordinates of all of the n_A members of category A:

$$x_{Am} = \frac{1}{n_A} \sum_{j \in A} x_{jm}, \quad (4)$$

and likewise for category B. According to the MPM, the probability that stimulus i is classified into category A is formalized as:

$$P(A \mid i) = \frac{sim(i, A)}{sim(i, A) + sim(i, B)}, \quad (5)$$

where $sim(i, A)$ denotes the similarity between stimulus i and the category A prototype, given by

$$sim(i, A) = \exp\left(-c\left[\sum_{m=1}^{M} w_m \mid x_{im} - x_{Am} \mid^r\right]^{\frac{1}{r}}\right). \quad (6)$$

Appendix C: Detailed PSP procedure and output

For each of the four category structures used in Nosofsky (1986), we ran the PSP algorithm on both the GCM and MPM. The range of the parameter space of both models was [0, 1] for the attention parameter w, and [0, 20] for the sensitivity parameter c, based on what is typically observed in experiments. A data pattern was defined as the ranking of the category A response percentages (rounded to the nearest integer) for the 16 stimuli. All stimuli with equal rankings (i.e., same category A response percentages) were assigned a midrank value (i.e., rank position equal to the average of all the tied positions). As suggested by Pitt et al. (2006), we ran PSP five times for each model and structure to ascertain consistency. The PSP output comprised of all data patterns generated across the runs, with the volume of each data pattern determined by calculating the average of its five assigned volumes (if the data pattern had not been generated in a certain run, it was assigned a volume of 0 %). This resulted in 8426, 14,848, 3902, and 6702 data patterns for the GCM, and 2142, 4444, 6222, and 5023 data patterns for the MPM in the criss-cross, interior-exterior, diagonal, and dimensional structure, respectively. The data patterns occupying at least one percent of the parameter space of the GCM and MPM can be found in the Supplementary Material, available on https://osf.io/qe5kc.

Appendix D: Recovery results sensitivity analysis

This section contains the results of the sensitivity analysis to evaluate the robustness of the PSP fits in the application example. Table 3 shows the PSP fits of the GCM and MPM in the four category learning structures obtained across the different data pattern definitions, considered stimuli and parameter ranges. The PSP fits between brackets reflect fits of models that appeared to be indistinguishable based on a visual inspection of their largest data patterns (the figures showing these data patterns in the diagonal structure can be found in the Supplementary Material, available on https://osf.io/qe5kc).

Table 3 PSP fits for the GCM and the MPM in the four category learning structures from Nosofsky (1986) across two different data pattern definitions (ranking and nominal), two sets of stimuli (all stimuli and the unassigned stimuli only), and three different parameter ranges of the sensitivity parameter c ([0, 20], [0, 10] and [0, 5])

		Ranking pattern				Nominal pattern			
		All stimuli		Unassigned stimuli		All stimuli		Unassigned stimuli	
Range c	Structure	GCM	MPM	GCM	MPM	GCM	MPM	GCM	MPM
[0, 20]	Criss-cross	0.17	0.59	0.29	0.61	0.01	0.50	0.01	0.50
	Interior-exterior	0.36	0.65	0.42	0.58	0.22	0.56	0.28	0.50
	Diagonal	0.13	0.30	(0.11)	(0.19)	0.05	0.20	(0.09)	(0.15)
	Dimensional	(0.19)	(0.26)	(0.21)	(0.27)	(0.002)	(0.05)	(0.002)	(0.06)
[0, 10]	Criss-cross	0.15	0.59	0.31	0.60	0.03	0.50	0.02	0.50
	Interior-exterior	0.36	0.65	0.42	0.58	0.24	0.56	0.30	0.50
	Diagonal	0.11	0.26	(0.10)	(0.15)	0.06	0.20	(0.08)	(0.15)
	Dimensional	(0.17)	(0.24)	(0.17)	(0.24)	(0.004)	(0.05)	(0.004)	(0.07)
[0, 5]	Criss-cross	0.14	0.57	0.33	0.59	0.04	0.50	0.03	0.50
	Interior-exterior	0.33	0.64	0.40	0.57	0.27	0.56	0.34	0.50
	Diagonal	0.09	0.23	(0.07)	(0.10)	0.08	0.20	(0.08)	(0.15)
	Dimensional	(0.15)	(0.20)	(0.14)	(0.20)	(0.01)	(0.05)	(0.007)	(0.06)

PSP fits between brackets reflect fits of indistinguishable models

References

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372–400.

Diaconis, P., & Graham, R. L. (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, 262–268.

Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., & Vee, E. (2006). Comparing partial rankings. *SIAM Journal on Discrete Mathematics*, *20*, 628–648.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (Eds.) (1996). *Markov chain Monte Carlo in practice*. Suffolk, UK: Chapman and Hall.

Hawkins, G. E., Brown, S. D., Steyvers, M., & Wagenmakers, E.-J. (2012). An optimal adjustment procedure to minimize experiment time in decisions with multiple alternatives. *Psychonomic Bulletin & Review*, *19*, 339–348.

Jeffreys, H. (1961). *Theory of probability*. Oxford: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*, 81–93.

Kendall, M. G. (1975). *Rank correlation methods*, 4th edn. London: Charles Griffin.

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

Lee, M., & Vanpaemel, W. (2016). *Determining informative priors for cognitive models*.

Marden, J. I. (1995). *Analyzing and modeling rank data*. London: Chapman & Hall.

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McFall, R. M., & Townsend, J. T. (1998). Foundations of psychological assessment: Implications for cognitive assessment in clinical science. *Psychological Assessment*, *10*, 316–330.

Minda, J. P., & Smith, J. .D. (2011). Prototype models of categorization: Basic formulation, predictions, and limitations. In Pothos, E., & Wills, A. (Eds.) *Formal approaches in categorization*. Cambridge, UK: Cambridge University Press.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*, 47–84.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*, 299–325.

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

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87–108.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*, 421–425.

Pitt, M. A., Myung, J. I., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

Pitt, M. A., Kim, W., & Myung, J. I. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin & Review*, *10*, 29–44.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*, 57–83.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 393–407.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*(13), 517.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289–310.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101.

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2013). A comparison of reinforcement learning models for the Iowa gambling task using parameter space partitioning. *The Journal of Problem Solving*, *5*, 2.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Vanpaemel, W. (2016). Prototypes, exemplars and the response scaling parameter: A Bayes factor perspective. *Journal of Mathematical Psychology*, 183–190.

Vanpaemel, W., & Lee, M. D. (2012a). The Bayesian evaluation of categorization models: Comment on Wills and Pothos (2012). *Psychological Bulletin*, *138*, 1253–1258.

Vanpaemel, W., & Lee, M. D. (2012b). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*, 1047–1056.

Vanpaemel, W., & Storms, G. (2010). Abstraction and model evaluation in category learning. *Behavior Research Methods*, *42*, 421–437.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*, 228–243.

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

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, *138*, 102–125.