# A Study of Individual Differences in Categorization with Redundancy

Farzin Shamloo, Sébastien Hélie

Purdue University

## Author Note

Farzin Shamloo, Psychological Sciences Department, Purdue University.

Sébastien Hélie, Psychological Sciences Department, Purdue University.

This research was funded in part by award #1349677-BCS from the National Science Foundation and award #2R01MH063760 from the National Institute of Mental Health. The authors would like to thank Madison Fansher, Sarina Farhadi, Julie Taylor, and Giovana Teles for their help with data collection.

Correspondence concerning this sample paper should be addressed to Farzin Shamloo. E-mail: farzin.shamloo@gmail.com

# Abstract

Humans and other animals are constantly learning new categories and making categorization decisions in their everyday life. However, different individuals may focus on different information when learning categories, which can impact the category representation and the information that is used when making categorization decisions. This article used computational modeling of behavioral data to take a closer look at this possibility in the context of a categorization task with redundancy. Iterative decision bound modeling and drift diffusion models were used to detect individual differences in human categorization performance. The results show that participants differ in terms of what stimulus features they learned and how they use the learned features. For example, while some participants only learn one stimulus dimension (which is sufficient for perfect accuracy), others learn both stimulus dimensions (which is not required for perfect accuracy). Among participants that learned both dimensions, some used both dimensions, while others show error and RT patterns suggesting the use of only one of the dimensions. The diversity of obtained results is problematic for existing categorization models and suggests that each categorization model may be able to account for the performance of some but not all participants.

**Keywords:** Category learning; Individual differences; Redundancy; Iterative decision bound modeling; Drift diffusion modeling.

# Introduction

Humans and other animals are constantly categorizing things that are encountered in everyday life. For example, people categorize other people they see in the street (stranger, friend), the food they eat (healthy, unhealthy), etc. The objects that people encounter everyday can typically be decomposed into numerous component dimensions (features). Importantly, most categorizations include some form of feature redundancy (e.g., Bahrick & Lickliter, 2000). As a result, there is often no need to learn all the features in order to successfully categorize objects. However, learning features that are not necessary for categorizing objects in a given task may help to better understand category characteristics, and the information that is redundant in one categorization task may become decisive in a future categorization tasks (e.g., Hélie et al., 2017). The goal of this study is to better understand the learning behavior of participants in a two-dimensional categorization task with redundancy where learning either one of the two dimensions by itself is sufficient to produce perfect accuracy. More specifically, the goal is to identify: 1) what is learned and 2) how responses are generated under redundancy. The results of the experiment are contextualized in relation to category learning theories and models.

## Hypothesis testing vs associative learning

Existing categorization models are in disagreement with how participants behave in a categorization task with redundancy. For example, some categorization models postulate that participants begin their learning process by testing simple unidimensional rules (e.g., COVIS; Ashby et al., 1998 and RULEX; Nosofsky et al., 1994). According to these models, if one of the tested simple rules is sufficient to satisfy task requirement, participants would not change their strategy and learning is complete. Therefore, such hypothesis-testing algorithms would predict

that in a two-dimensional categorization task where any of the two dimensions by itself suffice for perfect accuracy, a participant would begin by testing a unidimensional rule on one of the two dimensions, and since the rule would work, they would continue using that unidimensional rule.

On the other hand, many associative learning models (e.g., ALCOVE; Kruschke, 1992; Context Theory; Medin & Schaffer, 1978; GCM; Nosofsky, 1986) suggest that the probability of a stimulus being assigned to a category depends on the distance between that stimulus and previously stored exemplars from that category in perceptual space. Therefore, in a two-dimensional categorization task, associations would be made between category labels and both dimensions (assuming that previously stored exemplars contain both dimensions). However, most associative models (e.g., ALCOVE; Kruschke, 1992) include attentional restrictions (e.g., weights), which limits the number of dimensions being learned and/or the amount of attention allocated to each dimension. Therefore, the specifics of which dimensions are learned under redundancy would depend on how attentional weights are initialized, how they change, and the overall attentional capacity of the learner. For example, an associative model that initially assigns all attentional resources to one dimension and changes attentional weights only if an error feedback is received would produce predictions similar to RULEX in terms of what features are learned under redundancy. However, this same model would behave differently when attentional weights are initialized by assigning equal weight to all dimensions and change attentional weights with mechanisms other than relying only on error feedback. So while associative learning models are more flexible, it is unclear if this additional flexibility is needed to account for human category learning. The proposed redundancy protocol is a good starting point to begin testing these predictions.

## Cue validity vs category validity

Most category learning models implicitly assume that the goal of category learning is minimizing classification errors. For example, according to RULEX (Nosofsky et al., 1994), once a rule works, the strategy remains stable. Similarly, ALCOVE (Kruschke, 1992) updates its attentional weights using error feedback to allocate attention to the diagnostic features. In other words, these models presume that a dimension's importance is determined by cue validity (i.e., the probability of an object belonging to a category given a dimension's value). This approach is in contrast with an approach that presumes the primacy of category validity (i.e., the probability of having a feature value given the category) and some researchers (e.g., Ell et al., 2017, 2020; Markman & Ross, 2003) have argued that the primacy of cue validity over category validity in categorization models is caused by an over reliance of laboratory experiments on classification tasks. Outside the laboratory, features may be learned not necessarily to perform classification, but instead to later predict features of an item that belongs to a category (Anderson, 1990). For example, the end goal of categorizing an animal as a 'cat' might be to infer that it can chase mice – not the category label assignment itself (Hélie & Ashby, 2012). This suggests that in naturalistic category learning, even when the immediate task is to classify objects, there may be a tendency to learn about categories themselves and therefore a tendency to learn non-diagnostic features (e.g., Ell et al., 2017, 2020; Hélie, et al., 2017) in addition to diagnostic features (which are necessary in order to successfully categorize objects).

A possible way of mediating the tension between learning driven by cue-validity and learning driven by category validity is to distinguish the knowledge that is used to produce a response from the knowledge that is acquired not to produce the classification responses per se, but instead for other reasons (e.g., making inferences about an object's feature given the category

label). In other words, cue-validity can drive learning the dimensions that differentiate objects in a given classification task, but it may not be the sole driver in learning an objects' dimensions: Category-validity can promote the learning of other dimensions (e.g., by passive observation) that do not necessarily improve classification accuracy but instead improve feature inference for a given category. Note that in a two-dimensional categorization task with redundancy (the focus of this work), neither dimension is non-diagnostic. However, since it is possible for a participant to use only one of the dimensions to perform the task, it is still possible to distinguish between knowledge that is used to produce a response and knowledge that is latently learned (i.e., learned but not used). The proposed new analysis methods included in this article aim at being able to disentangle these two types of knowledge at the individual participant level.

## The Present study

This study focused on a categorization task with redundancy and used behavioral measures (accuracy and response time) to determine what is learned and what is used in a redundant categorization task. The experiment was composed of two phases. In the training phase, participants were asked to perform a two-alternative forced choice task where participants received feedback after each trial. There were four categories in the perceptual space overall, but in the training phase participants only performed two (out of the six total possible) comparisons. In both cases, using any of the dimensions by itself was enough for perfect classification accuracy. In the test phase, participants no longer received feedback and were tested on all possible two-alternative forced choice comparisons. Three different analyses were performed: Learned Knowledge, iDBM Analysis, and DDM Analysis. The goal of all three analyses was to investigate what information is learned and used under redundancy (i.e., the training phase). The first analysis (Learned Knowledge) used the data from the test phase to determine what was learned during the training phase. The iDBM Analysis (Hélie et al., 2017) tested whether

participants explored new rules even when an old rule worked (a common assumption in hypothesis-testing models). The DDM Analysis used drift diffusion models (Ratcliff, 1978; Ratcliff & McKoon, 2008; Ashby, 2000) to determine which dimensions were used by each participant when redundancy is present. Specifically, it posited that different strategies impose different difficulty maps on the perceptual space, and assigned each participant to a response strategy by comparing the goodness of fit of different difficulty maps to each participant's performance.

Lastly, the results from the three analyses were put together to show the existence of participants who do not follow the learning patterns suggested by popular categorization model assumptions. For example, the existence of rule switchers (assessed in the iDBM analysis) challenges the assumption that changing rule does not occur if a unidimensional rule works. The existence of participants who learn and use both dimensions challenges the assumption that attentional weights are initially assigned to one dimension and change only if needed to enhance categorization performance (cue validity). On the other hand, the existence of participants who learn and use only one dimension challenges the generality of the primacy of category validity in classification tasks. Finally, the existence of participants who learn both dimensions but show error and RT patterns suggesting the use of only one of the dimensions challenges that learning mechanisms rely only on error feedback and suggest that information can be learned latently. The diversity of results obtained suggests that each categorization model may be able to account for the performance of some but not all participants.

# Experiment

## Method

The goal of this experiment was to study category learning under redundancy using two-dimensional stimuli. The experiment consisted of six blocks and in the first five blocks perfect accuracy could be achieved by using either one of the two dimensions alone (the redundant trials). The sixth (last) block of the experiment tested participants' knowledge on each of the dimensions and was used to determine the learned knowledge of each participant.

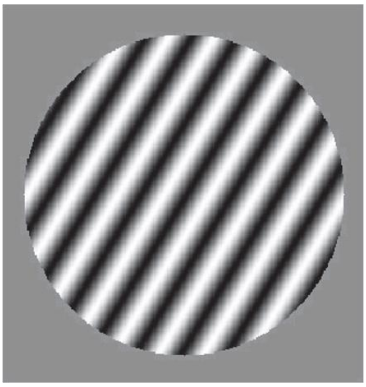
### Participants

One hundred seventy Purdue University undergraduate students participated in the study and received partial credit to fulfill a course requirement.

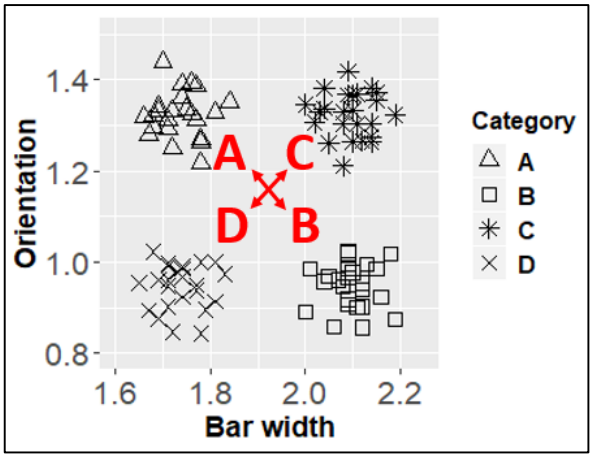
### Material

The stimuli were sine-wave gratings of constant contrast and size that differed in frequency (ranging from 1.65 to 2.21 cycles per degree) and orientation (counterclockwise rotation from horizontal ranging from 0.82 to 1.44 radians). Stimulus presentation and response recording was done using the Psychophysics Toolbox in MATLAB (Brainard, 1997). There were four categories (arbitrarily labeled "A", "B", "C" and "D") and in each trial participants were shown a stimulus and asked to choose between two of the categories. Stimuli were generated using bivariate normal distributions with the following parameters: $\mu_A = (1.736, 1.322), \mu_B = (2.096, 0.945), \mu_C = (2.096, 1.322), \mu_D = (1.736, 0.945), \sum_A = \sum_B = \sum_C = \sum_D = \begin{pmatrix} 0.002 & 0 \\ 0 & 0.002 \end{pmatrix}$. Ninety-six stimuli were generated (twenty-four from each category) which were shuffled in the beginning of each of the training blocks. One hundred forty-four stimuli (thirty-six of each category) were generated for the test phase. Figure1a shows a sample stimulus and Figures 1b and 1c show the generated stimuli in the training and test phase, respectively. Participants only performed two types of categorization trials in the training phase. In each trial

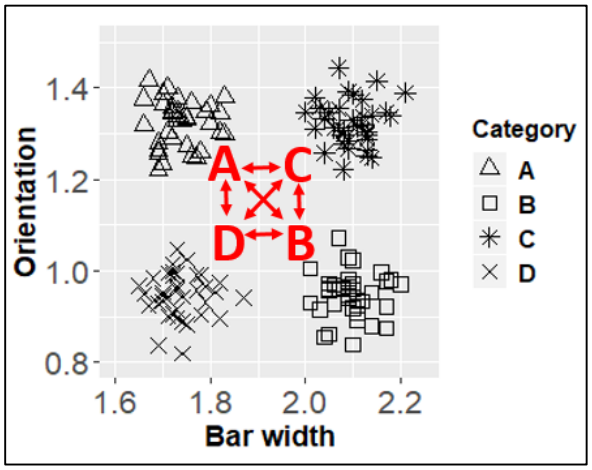
the question shown on the screen was either “A or B?” or “C or D?”, and note that in both of “A or B?” and “C or D?” knowledge on any one of the two dimensions is enough to distinguish between the two categories. In the test phase, participants performed all possible two choice categorizations (“A or B?”, “A or C?”, “A or D?”, “B or C?”, “B or D?”, “C or D?”). Red arrows in Figures 1b and 1c show the categories that were compared together in each phase of the experiment. Participants responded using a standard keyboard and in all of the trials, “d”, “k”, “x” and “m” keys were used to choose categories “A”, “B”, “C” and “D” respectively.



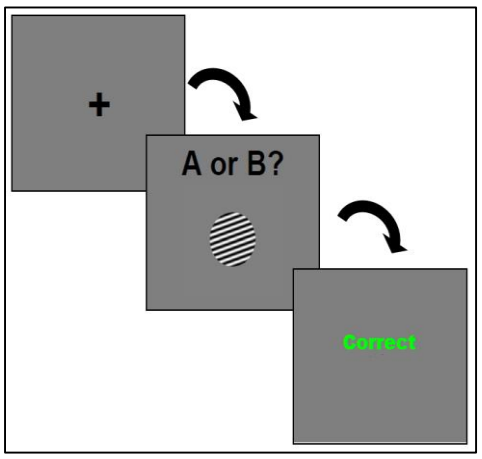
(a)




(b)




(c)




(d)

*Figure 1.* (a) An example stimulus (b) The stimuli used in the training phase. Red arrows indicate the comparisons participants were asked to do (c) The stimuli used in the test phase. Red arrows indicate the comparisons participants were asked to do (d) An example of a trial sequence in the training phase of the experiment. No feedback were given in the test phase.

## Procedure

Participants were told that they would be participating in a two choice categorization task where stimuli are sine-wave gratings that differed in bar width and orientation. They were told that there are four categories, and in each trial, a question on top of the screen would ask them to choose between two of the four categories. The experiment was divided into six blocks and participants categorized ninety-six stimuli in each of the first five blocks and 144 stimuli in the sixth block. Participants were told that during the first five blocks, they would receive feedback but no feedback would be given in the last block. Each training trial started with a fixation cross that was presented at the center of the screen for 1500 ms. Then the fixation cross was replaced by the stimulus and categorization question. As soon as the participant responded, the stimulus and question were replaced by feedback (green “Correct” for correct responses and red “Incorrect” for incorrect responses) which stayed on the screen for 750 ms. In trials where the participant did not respond within five seconds, the words “Too Slow!” was shown in black in the center of the screen. The timed-out trials were counted as errors. Test trials followed the same sequence except that no feedback was given to participants. Figure 1d shows the display sequence for a trial in the training phase.

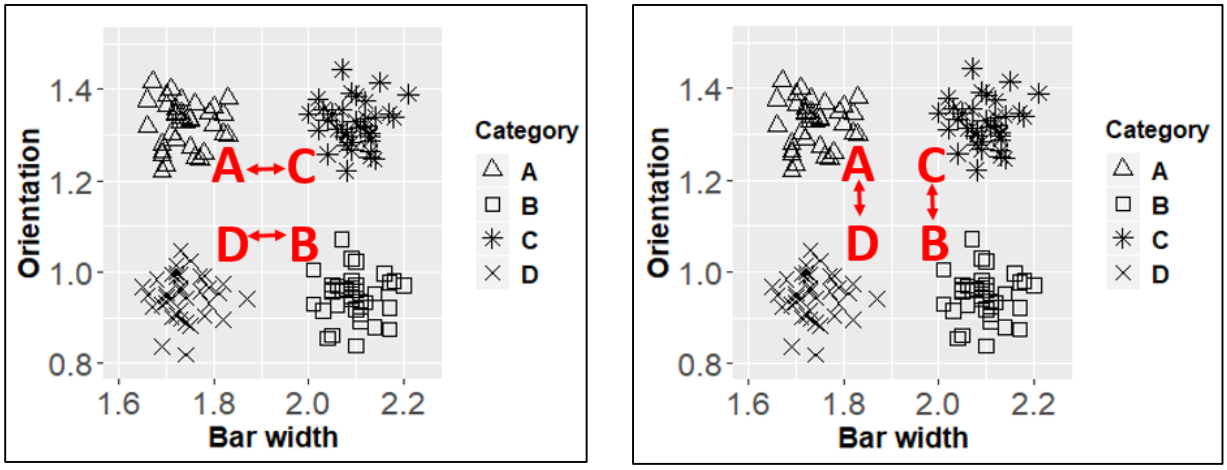
The remainder of this article is divided into three sections: “Learned Knowledge”, “iDBM Analysis” and “DDM Analysis”. We first identified the learned knowledge of each participant by analyzing the test phase (in the “Learned Knowledge” section). Next, the iDBM

and DDM analyses focused on the training data to identify response patterns under redundancy of information. The iDBM Analysis used a hypothesis-testing framework to specifically test whether participants switched rules, while the DDM Analysis used a more neutral framework to identify participants' strategies.

# Results

## Learned knowledge

Figure 2 shows two types of categorization tasks participants performed in the test phase. Figure 2a shows trials in which knowledge on bar width was decisive ("BW" trials) and Figure 2b shows trials in which knowledge on orientation was decisive ("OR" trials).




(a) (b)

*Figure 2.* Two types of categorization trials that participants did in the test phase (a) "BW" trials: "A or C?" and "B or D?" trials, where knowledge on bar width differences is necessary. (b) "OR" trials: "A or D?" and "B or C?" trials, where knowledge on orientation differences is necessary.

Participants received no feedback in the test phase and therefore, it is safe to assume that success on "BW" trials meant that knowledge on bar width was acquired during training and

similarly, success on "OR" trials meant that knowledge on orientation was acquired during the training. Participants were divided into five groups based on their learned knowledge. The five groups correspond to participants that acquired knowledge on both dimensions ("Learned Both"), participants that acquired knowledge only on bar width ("Learned BW"), participants that acquired knowledge only on orientation ("Learned OR"), participants that acquired knowledge on none of the dimensions ("Non-Learner"), and participants for whom the data is not conclusive and therefore do not fit to any of the other four well-defined groups ("Unclear"). Participants' responses for each of "BW" and "OR" trials were modeled using a binomial distribution with success probability ($p$)of $\Theta_{BW}$ and $\Theta_{OR}$ respectively, and n=48 number of trials. For each of "BW" and "OR" trials two alternatives are considered: $\Theta_{BW} = 0.5$ and $\Theta_{BW} > 0.5$ for the "BW" trials and similarly, $\Theta_{OR} = 0.5$ and $\Theta_{OR} > 0.5$ for the "OR" trials. Bayes factor for each of "BW" and "OR" trials were calculated in the following way:

$$BF = \frac{p(D|M_2)}{p(D|M_1)}$$

$$M_1\colon\ \theta = 0.5\ and\ M_2\colon\ \theta > 0.5$$

$$p(D|M_1) = \binom{48}{k} \times 0.5^{48}$$

$$Where\ "k"\ is\ number\ of\ correct\ responses$$

$$p(D|M_2) = \int p(D|\,\theta, M_2)p(\theta|M_2)d\theta$$

The last block of training was used to choose a proper $p(\theta|M_2)$. Using the "fitdist" function in R (Delignette-Muller & Dutang, 2015) and comparing different distributions (Normal, Lognormal, Weibull, Logistic and Gamma), a Weibull with a shape parameter of 16.9 and scale parameter of 0.94 turned out to be the best fit.

Having chosen $p(\theta|M_2)$, each participants' Bayes factor of "BW" and "OR" trials ($BF_{BW}$ and $BF_{OR}$ respectively) was calculated using the number of correct responses ("$k$" in the formula for $BF$) in each of "BW" and "OR" and each participant was assigned to one of the five mentioned groups ("Learned Both", "Learned BW", "Learned OR", "Non-Learner" and "Unclear"). A Bayes factor larger than 3 was considered strong evidence for $\theta > 0.5$ model and a Bayes factor smaller than 0.33 was considered strong evidence for the $\theta = 0.5$ model [thresholds on the Bayes factors are based on suggested values in Jeffreys (1998)]. The values of Bayes factor between 0.33 and 3 were considered non-conclusive and if any of $BF_{BW}$ or $BF_{OR}$ fell between 0.33 and 3, the participant was assigned to the "Unclear" group.

Figure 3 summarizes the test performance and assigned learned knowledge. Each circle represents a participant (the number inside the circle is the participant number) and the x-axis and y-axis are participant's accuracy in "BW" and "OR" trials respectively. Note that as previously stated, no feedback was given in the test phase, and it was assumed that the knowledge learned by each participant was acquired during the training phase (i.e., under redundancy)[1]. Figure 3 confirms that there are individual differences in the learned knowledge when there is redundancy. Some participants acquire knowledge on both dimensions (northeast of the plot) while some participants only acquire knowledge on one of the dimensions (located

[1] Note that this is not necessarily true for all category structures and learning paradigms. There is evidence for unsupervised category learning (e.g., Ashby et al., 1999). However, as shown by Ell and Ashby (2012), unsupervised learning depends on factors such as within-category variability and between-category distance. Here the assumption of no learning in the test phase was made for three reasons. First, the class separation index of the category structure studied in this article ($d' = 8.05$ for each of BW and OR trials) is close to those for which Ell and Ashby (2012) showed little unsupervised learning. Second, and more importantly, Ell and Ashby (2012) studied unsupervised learning alternating between observation only and response blocks (overall ten blocks, each with eighty trials, forty of each category) while in the test phase of this study, there was no observation only blocks and only one response block of one hundred forty-four trials (thirty-six of each category). Third, while there were only two categories in Ell and Ashby (2012) and one comparison throughout the blocks, the test phase of this study had four categories and six different comparisons. Hence, the conditions for unsupervised learning observed in Ell & Ashby (2012) were not present in this experiment's test block.

either on the northwest of the plot or on the southeast). Note that the assigned labels base on the Bayes factor intuitively make sense: Participants with high average accuracy on both "BW" and "OR" trials are assigned to "Learned Both" group, participants with high accuracy on "BW" trials and low accuracy on "OR" trials are assigned to "Learned BW" group, participants with low accuracy on "BW" trials and high accuracy on "OR" trials are assigned to "Learned OR" group, participants with low accuracy on both of "BW" trials and "OR" trials are assigned to "Non-Learner" group and participants with accuracy of around 65% on one or both of "BW" or "OR" trials are assigned to "Unclear", since it is not clear whether they learned the dimension with 65% accuracy or not. There were thirteen "Non-Learner" participants who were excluded from the remaining analyses.

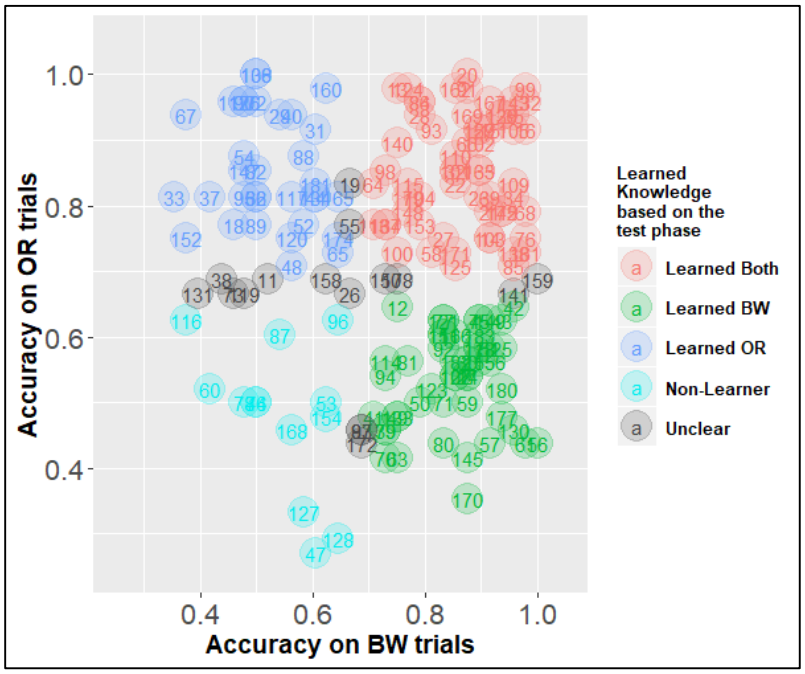



*Figure 3.* Test performance of participants. The x-axis is the mean accuracy on trials where knowledge on bar width was necessary to categorize the stimulus and the y-axis is the mean accuracy on trials where knowledge on orientation was necessary to categorize the stimulus.

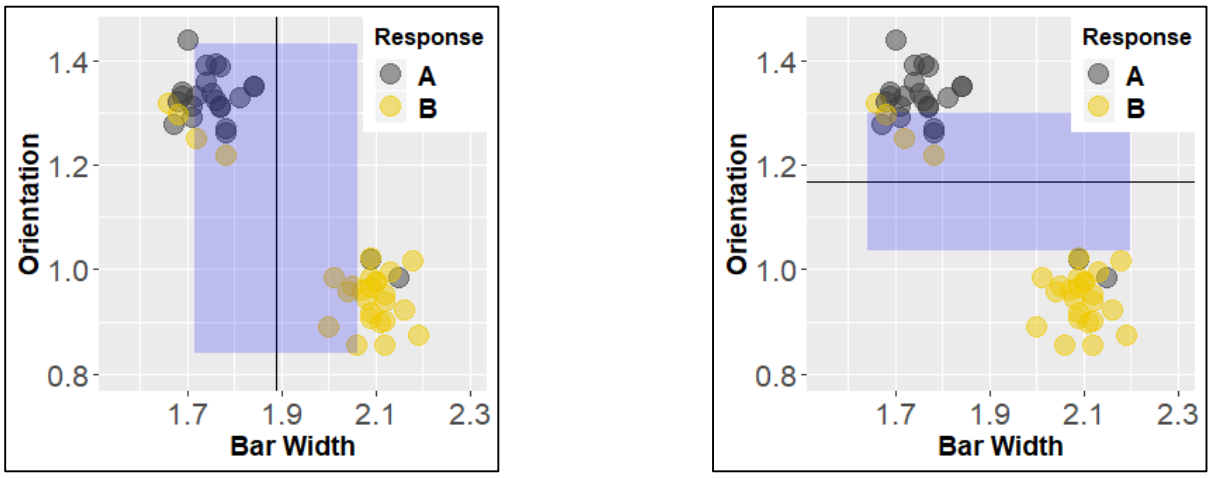
Each circle is a participant and color of each participant shows whether they learned both dimensions, only bar width, only orientation, neither, or unclear.

In the next two analyses, response patterns of participants in the training phase are directly analyzed. The iDBM Analysis tests whether participants switch rules and the DDM Analysis determines the difficulty map that best explains each participant's response strategy. The results of this section that determined the learned knowledge of participants is put together with the result of the DDM Analysis in the Discussion section and implications for the distinction between the learned and used knowledge is discussed.

## iDBM Analysis

General recognition theory (Ashby & Townsend, 1986) is an extension of signal detection theory in multidimensional spaces that has been used in numerous contexts in the past 30 years (e.g., Ashby & Gott, 1988; Ashby & Perrin, 1988; Maddox et al., 2002). When GRT is used to model participants' response patterns in a categorization task, it is often called a decision bound model (Ashby & Soto, 2015). Decision bound models assume that participants divide perceptual space using bounds and use these bounds to make categorization decisions. iDBM Analysis uses iterative Decision Bound Modeling (iDBM; Hélie et al., 2017) to identify the rules that each participant used. iDBM fits decision bound models iteratively to data and outputs the best fitting decision bound on each trial based in decisions in a moving time window. iDBM was fit to "A or B?" and "C or D?" trials separately. Figure 4 is an illustration of what iDBM does in one of its iterations. In Figure 4 the best fitting bound on each dimension is fitted to trials 10-150 of one of the participants and the maximum likelihood values are compared. In the instance shown in Figure 4, the participant seemed to be using the orientation dimension, since the errors are close to the bound on orientation, which is reflected in the likelihood values: -16.19 for the bound on orientation and -23.62 for the bound fitted on the bar width dimension (more details in

Hélie et al., 2017). The iDBM was set to output one of the following three models for each trial: guessing model, unidimensional bound on bar width, or unidimensional bound on orientation.




(a) (b)

*Figure 4.* A visualization of how iDBM works. The bounds are fitted to trials 10-150 of participant 109 (a) The bound fitted on bar width (b) The bound fitted on orientation.

For each of "A or B?" and "C or D?" tasks, participants started by guessing and there are four possibilities for the rest of the experiment: continue guessing untill the end of the experiment, using a rule on bar width for the rest of the experiment, using a rule on orientation for the rest of the experiment, and switching between rules. Therefore there are overall sixteen possibilities for a participant's iDBM label. The iDBM labels are set by concatenating rule usage on "A or B?" and "C or D?" tasks. For example a participant that is using a rule on bar width on "A or B?" task but switches between rules in "C or D?" task is labeled as BW_Switch and a participant that uses a rule on orientation on both "A or B?" and "C or D?" tasks is labeled as OR_OR.

## iDBM Analysis results

Figure 5 shows the result of the iDBM Analysis. The labels extracted from the iDBM Analysis are displayed in the context of the results from previous section (the learned knowledge). In other words, each circle represents a participant, the x-axis and y-axis are accuracy on "BW" and "OR" trials of the test phase, and the iDBM assigned label of each participant is coded in the color of each circle. The nine different iDBM labels are color coded into five groups. Blue denotes participants that used a rule on bar width throughout the experiment in both "A or B?" and "C or D?" trials. Red denotes participants that used a rule on orientation throughout the experiment in both "A or B?" and "C or D?" trials. Green denotes participants that did not switch rules, but used different rules on "A or B?" and "C or D?" trials. Light gray denotes participants that switched rule on one of the two trial types and finally, dark gray denotes participants that switched rules on both "A or B?" and "C or D?" trials. The red, blue and green participant's rule usage pattern is compatible with the rule-based categorization models that predict participants begin by testing unidimensional rules and if the rule works, they do not switch. However, as Figure 5 shows, there are sixteen participants (colored light and dark gray) that switched rule in one or both types of trials.

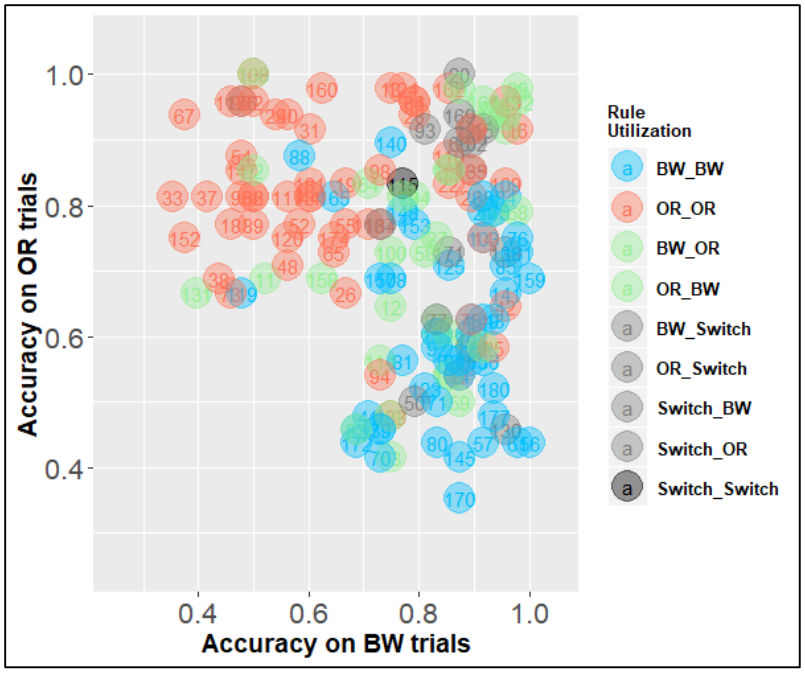



*Figure 5.* The rule usage of participants based on the iDBM Analysis. Each circle represents a participant. The color of the circles show the rule usage and its location shows the learned knowledge (x-axis is test accuracy on "BW" trials and y-axis is test accuracy on "OR" trials).

Table 1 summarizes the correspondence between labels assigned by iDBM and the learned knowledge for participants.

Table 1.

*The confusion table for the relation between identified used rule (based on iDBM) and learned knowledge. Participants with "Unclear" learned knowledge are not shown in the table and all the participants with one rule switch are grouped together as "1 Switch".*

| | | Learned knowledge | | |
|---|---|---|---|---|
| | | Learned BW | Learned OR | Learned Both |
| Rule usage | BW_BW | 26 | 2 | 12 |

| | | | | |
|---|---|---|---|---|
| | OR_OR | 6 | 28 | 19 |
| | BW_OR<br>OR_BW | 12 | 2 | 18 |
| | 1 Switch | 4 | 1 | 10 |
| | Switch_Switch | 0 | 0 | 1 |

The existence of participants that switched in one or both trial types challenges the assumption made by popular categorization models (e.g., RULEX, COVIS) that if a rule works participants do not switch rules. Another important finding is that, since the iDBM Analysis determined rule usage by using the location of errors in perceptual space, and some participants used a rule on a single dimension throughout the experiment (i.e., BW_BW and OR_OR group) but learned both dimensions, learning for at least some participants occurred by mechanisms other than only relying on error feedback. Having said that, the iDBM Analysis only used errors and did not consider RTs. Its rule-based assumption is also too restrictive. The following DDM Analysis used a broader framework not restricted to hypothesis testing or associative models.

## DDM Analysis

The iDBM Analysis suggests that assuming that all participants begin by testing unidimensional rules might be insufficient. The analysis shows that some participants continued using the same rule throughout the experiment (as predicted by many categorization models), but other participants started testing and using different unidimensional rules. While the iDBM Analysis was well suited to test the predictions of rule-based model components, the assumptions made were too restrictive. In the DDM Analysis, a more general method was used. It was assumed that different strategies produce different difficulty maps on a given set of trials. More

specifically, the difficulty of each trial was assumed to be a function of its location in the perceptual space with the specific shape of the function depending on the strategy that was used to produce a response. Drift Diffusion Models (DDM; Ratcliff, 1978; Ratcliff & McKoon, 2008) were used to model different difficulty maps. DDMs have been used to model two-alternative forced choice tasks in many different research areas (e.g., Aging: Ratcliff et al., 2004; Aphasia: Ratcliff et al., 2004). Ashby (2000) used DDM to study accuracy and response times in categorization tasks and introduced a dynamic version of the decision bound models called stochastic GRT. A DDM has a noisy evidence accumulator and two decision boundaries. Percept and accumulated evidence are stochastic processes that are usually modeled as a discrete random walk process. Figure 6 shows an accumulator with visualization of the process for three trials of a choice task with two options.

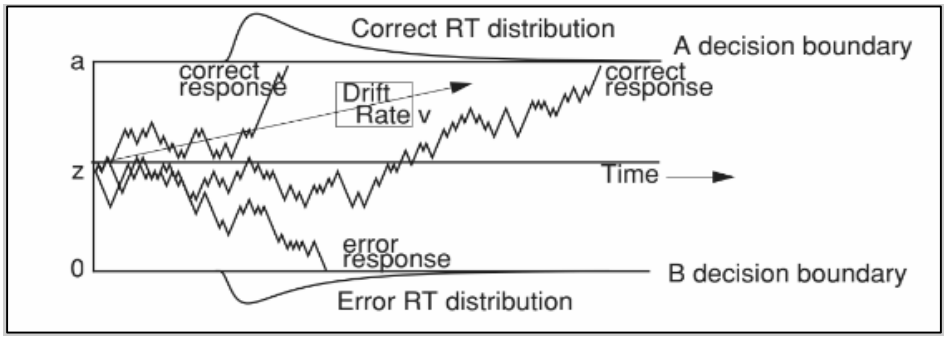
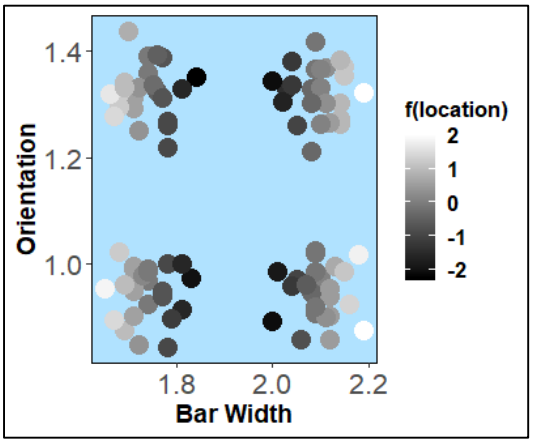



*Figure 6.* Three instances of a drift diffusion process. a, z, $v$, reflect speed accuracy trade off, bias and difficulty respectively. Figure is from Ratcliff and McKoon (2008).

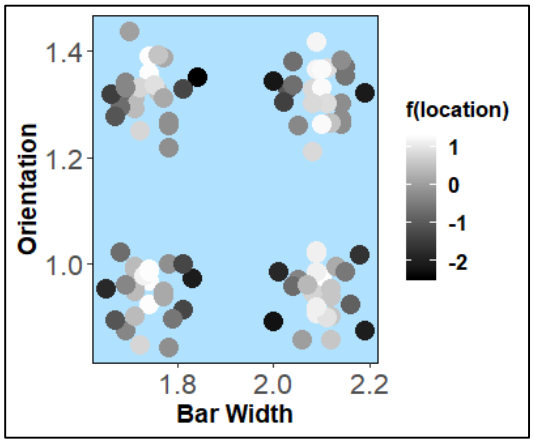
In DDMs difficulty is modeled by parameter $v$. In this work, we are assuming that different response strategies impose different trial by trial difficulty (Hélie et al., 2010). The following formulation was used to model various strategies:

$$v = v_0 + \beta \times f(location\ of\ stimulus)$$

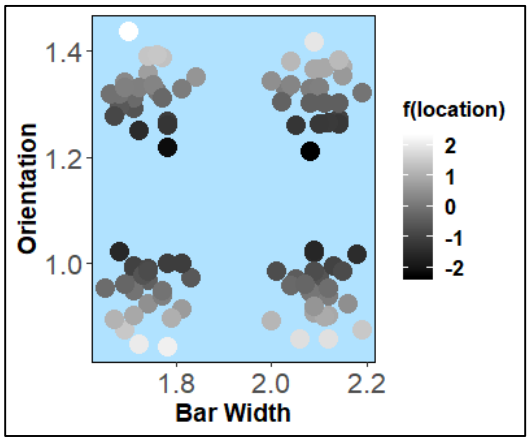
$f(location\ of\ stimulus)$ is a covariate which varies across strategies. Figure 7 shows four unidimensional strategies that were considered. The four unidimensional strategies differ in which dimension they use (bar width: 7a and 7b, orientation: 7c and 7d) and whether the difficulty was boundary based (i.e., most difficult trials being those closer to members of the other category: 7a and 7c) or typicality based (i.e., most difficult trials being those far from the center of the category: 7b and 7d).



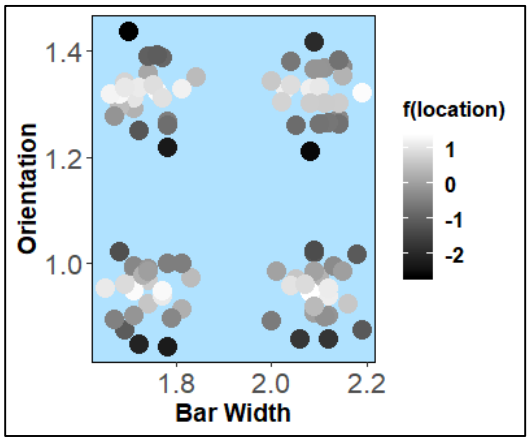

(a)




(b)



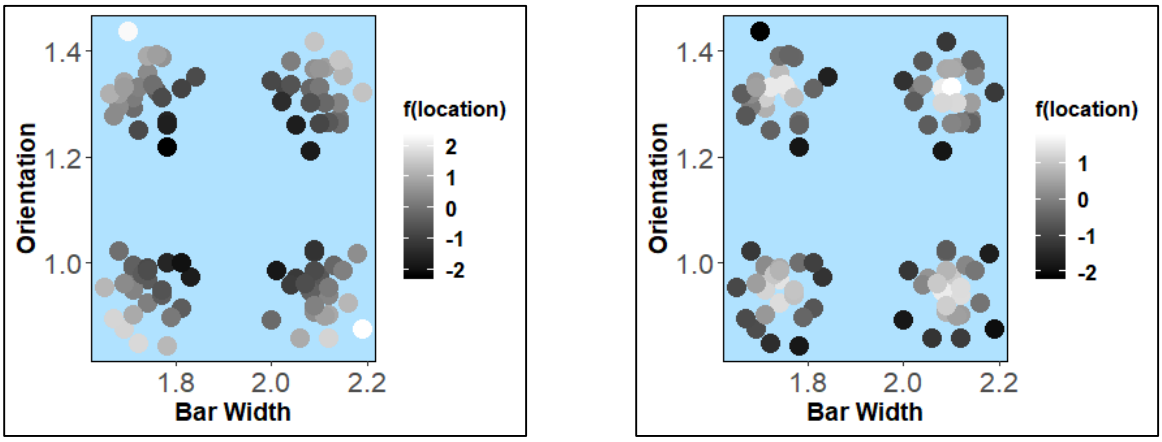

(c)




(d)

*Figure 7.* The covariate maps expected to fit best to participants with a unidimensional strategy. (a) and (b) Unidimensional strategies on bar width ("BW boundary" and "BW prototype"). (c) and (d) Unidimensional strategies on orientation ("OR boundary" and "OR prototype").

Figure 8 shows $f(location\ of\ stimulus)$ for the two-dimensional strategies that are considered. Figure 8a shows a boundary based two-dimensional strategy where trial difficulty increases further away from the members of the other category (labeled as "2D boundary") and Figure 8b shows a typicality based two-dimensional difficulty map where trial difficulty increases further away from the center of category(labeled as "2D prototype").




(a) (b)

*Figure 8.* The covariate maps expected to fit best to participants that used both dimensions. (a) "2D boundary" strategy (b) "2D prototype" strategy.

In addition to the six difficulty maps shown in Figures 7 and 8, we also considered a DDM with no covariate (i.e., $\beta = 0$),which is expected to fit best to participants for whom the relative difficulty of a trial does not depend on its location in the perceptual space (called "No covariate" model). The "No covariate" model is expected to capture response patterns of participants that have mastered the redundancy task well enough that they perceive all stimuli

equally easy. The six DDMs corresponding to the six difficulty maps in Figures 7 and 8 and the DDM with no covariate were fitted to the last three blocks of training of each participant separately and the best fitting model were identified using a model selection criterion (described in the next section). The HDDM (Wiecki et al., 2013) package was used to estimate parameters of the seven models for each participant. Even though DDMs were introduced over forty years ago (Ratcliff, 1978), there is still ongoing research on parameter estimation methods (for a review: Ratcliff & Childers, 2015). HDDM allows adding covariates to DDM parameters, which makes it possible to assess the effect of trial-by-trial variability on the parameters which makes it a fitting tool for the purposes of this study.

Originally, the trial-by-trial variability feature of HDDM was meant to incorporate neural data in estimating DDM parameters (Wiecki et al., 2013) and showing associations between neural measurements and changes in DDM parameters. For example, Cavanagh et al., (2011) showed that in high conflict trials of a reward-based decision-making task, trial-by-trial variability in frontal theta was associated with changes in the decision threshold parameter. In this study this feature was used to model the behavioral data not by using 'external' information (neural data) but by imposing a form of relation between objective features of stimuli and participant's perception in order to test whether the imposed form of relation can explain the RT and accuracy patterns. The imposed form of relations in this study was the difficulty maps, which differentiated between different strategies by considering different ways that the location of a stimulus in category space can affect $v$. However, this approach can be used to test any hypothesis that implies a relationship between stimulus location in category space and perceived difficulty, response threshold or bias.

### Model selection process in DDM Analysis

Two measures were used to assess the best fitting model, the DIC score and percentage of posterior samples of $\beta$ that are bigger than zero. DIC is a measure similar to AIC, which is used when the model fitting is done with Bayesian methods and posterior samples of parameters are available (Spiegelhalter et al., 2002). Similar to AIC, a DIC score is a goodness of fit measure that penalizes the number of parameters but unlike AIC, it is not possible to translate the scores to relative probabilities (i.e., the probability that a model provides the best description for the data among the considered models). However, it seems that the rule of thumb used in AIC works for DIC scores as well (Spiegelhalter et al., 2002), which is to consider all models that are within 1 point of the best model (i.e., the model with lowest DIC) to be relatively good. However, based on our simulation (see Appendix A) and in order to be conservative and avoid misidentifications, a model with the lowest DIC score was chosen only if its DIC score was at least five points better than the next best model. In addition to DIC scores, the posterior samples of $\beta$ parameter (except "No covariate" model) were also considered and a model was chosen only if 99% of its $\beta$ samples were greater than zero.

To summarize, a model is picked if (a) it has the smallest DIC score and its score is bigger than next best model by five points and (b) if its $\beta$ parameter is positive with a probability of 99% or higher. If none of the models satisfy the two mentioned conditions, then DDM Analysis does not assign a strategy to that participant. The strategy of such participants is labeled "None" in the following sections.

### DDM Analysis Results

Figure 9 shows the result of the DDM Analysis. The labels extracted from the DDM Analysis are again displayed in the context of the Learned knowledge section. In other words, each circle represents a participant, the x-axis and y-axis are accuracy on "BW" and "OR" trials

of the test phase and the strategy identified with the DDM for each participant is coded by the color of each circle. The DDM labels are color coded into four groups. Blue denotes participants that used a unidimensional strategy on bar width, green denotes participants that used a unidimensional strategy on orientation (dark and light shades for "OR boundary" and "OR prototype" respectively), red denotes participants that used a two-dimensional strategy (dark and light shades for "2D boundary" and "2D prototype" respectively) and finally, 'No Covariate' and 'None' labels are colored gray. The reason for grouping 'No Covariate' and 'None' models is that in the simulation (Appendix A), it was shown that in many cases data generated by other models was assigned to these two groups.

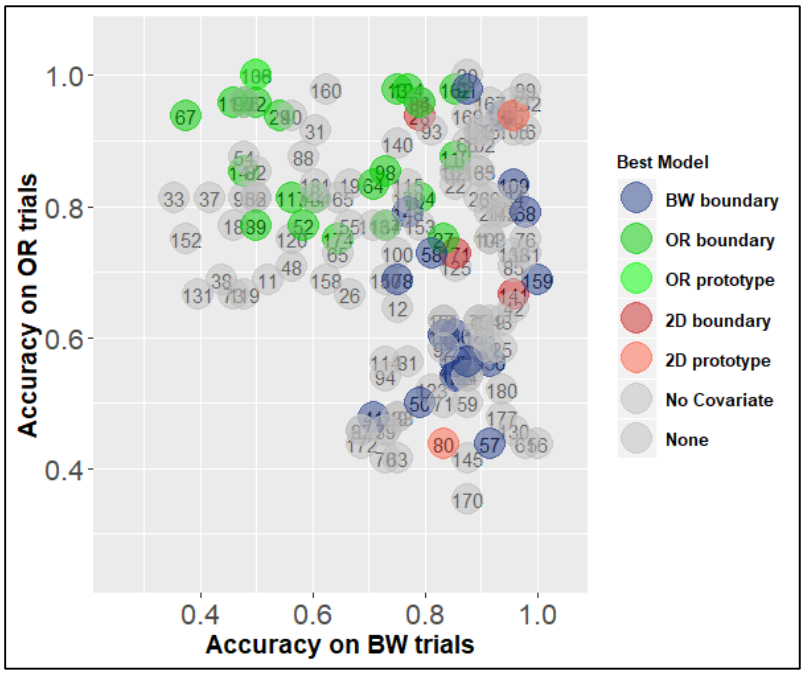



*Figure 9.* The response strategy identified for each participants based on the DDM Analysis. Each circle represents a participant. The color of the circle shows the strategy identified and its location shows the learned knowledge (x-axis is test accuracy on "BW" trials and y-axis is test accuracy on "OR" trials).

Table 2 summarizes the correspondence between labels assigned by DDM and the learned knowledge for participants.

Table 2.

*The confusion table for the relation between identified strategy (based on DDM) and learned knowledge. Participants with "Unclear" learning are not shown in the table.*

| | | Learned knowledge | | |
|---|---|---|---|---|
| | | Learned BW | Learned OR | Learned Both |
| DDM identified strategy | BW boundary | 11 | 0 | 6 |
| | OR boundary | 0 | 11 | 10 |
| | OR prototype | 0 | 1 | 0 |
| | 2D boundary | 0 | 0 | 2 |
| | 2D prototype | 1 | 0 | 1 |
| | No Covariate | 22 | 9 | 23 |
| | None | 14 | 12 | 18 |

The results of the DDM Analysis show that in addition to participants that learned only one dimension, it seems that 16 out of 19 participants that were not assigned to a vague model (i.e., No Covariate' and 'None') and learned both dimensions also relied on unidimensional strategies. This suggests that there is a dissociation between features that are learned and affected the response pattern and features that are learned but did not affect the response pattern.

# General Discussion

Many real world categorization situations include redundancy of information. For example, infants are exposed to much redundant information (Bahrick & Lickliter, 2000) and therefore, understanding the decision making process under redundancy may be critical to understanding real-world category learning. This article studied individual differences of participants in a two-dimensional category learning task with redundancy. The results were used to assess the limitations of existing categorization models. The article consisted of three analyses: Learned knowledge, iDBM Analysis and DDM Analysis.

The Learned Knowledge analysis shows that there are participants who only learn one stimulus dimension while other participants learn both stimulus dimensions, even though learning either dimension alone is sufficient for accurate classification. Individual differences in learned knowledge were shown by testing participants on new types of trials where knowledge of previously redundant dimensions became decisive. These individual differences suggest that there are various tendencies and depending on which tendency prevails, a participant can learn one or both of the stimulus dimensions under redundancy conditions. One tendency is to prioritize cue validity and spend the minimum amount of attentional resources required to do the classification task. Another tendency is prioritizing category validity and learning as much information as possible about the category features. This study does not determine why for some participants one of the tendencies prevail over the other, but it can potentially relate to individual differences in motivation or more general cognitive characteristics such as working memory capacity. After establishing which dimensions were learned by testing participants in the non-redundant phase of the experiment, the response patterns of the training phase where participants

performed two-dimensional categorizations with redundancy were analyzed to determine which dimension(s) were used. Determining the dimensions used by each participant demonstrated a potential dissociation between learning driven by cue-validity and learning driven by category-validity by showing that there are participants who learned both dimensions but only used one of them when making responses. In other words, this suggests that for such participants, while one of the dimensions was learned and used to make responses (learning driven by cue-validity), the other dimension was learned without being used (learning driven by category-validity). Two separate analyses (iDBM and DDM) were performed to determine how the learned dimension(s) were used.

The iDBM Analysis was used to identify strategy shifts during learning. However, this analysis made the restrictive assumption that participants used unidimensional rules, which limited the conclusions to rule-based models. The iDBM results showed that there are many participants that switch rules in the absence of errors, which challenges the assumption made by models that posit that rule switch does not happen if a rule works (e.g., COVIS; Ashby et al., 1998 and RULEX; Nosofsky et al., 1994). Even though the scope of the iDBM Analysis did not allow for making any general conclusions about the strategy used (only count strategy switch), it is worth mentioning that according to iDBM (which relies on the location of errors in the perceptual space) many participants whose error pattern suggested that they only used one dimension throughout the experiment learned both dimensions. This suggests that participants may learn by mechanisms other than relying solely on error feedback and also suggests a dissociation between what information was learned by the participants and what information was being used to produce responses. However, since iDBM only considered rule-based response strategies, it was not possible to make any strong conclusions about the response strategy being

used. A more general framework was used in the DDM Analysis to identify the participants' response strategies.

The DDM Analysis assumed that response strategies differ in the way a participant perceives the relative difficulty of individual stimuli. There is no limit to the number of difficulty maps that could be considered, but we chose four maps corresponding to unidimensional strategies and two maps corresponding to two-dimensional strategies. Half of the maps were boundary-based (i.e., easy stimuli are far from members of other categories) and half were typically based (i.e., the closer the stimulus is to the center of the category, the easier the trial). The DDM Analysis showed that some participants used unidimensional strategies while learning both dimensions, which suggests a dissociation between learned and used knowledge. It also further support earlier iDBM findings that not all learning is based on feedback errors and participants may have only used one dimension while latently learning the other dimension (Tolman, 1948; Tolman & Honzik, 1930).

There are at least two possible mechanisms to explain how latent learning occurs that differ in the timing of attentional allocation. One possibility is that participants start by attending to only one dimension and only pay attention to other dimension when the task becomes easy enough for them to have available attentional resources. The second possibility is that participants start by attending to both dimensions but decide to abstract a simple unidimensional rule and execute the response according to the rule while latently learning the other features. There is evidence supporting the second possibility in Rehder and Hoffman (2005) where eye-tracking data showed that most participants fixated on all dimensions in the beginning of an experiment while behavioral measures suggested many of the same participants were testing unidimensional rules. This seemingly contradictory evidence was explained by suggesting that

participants are “opportunistic” learners who take advantage of more than one learning mechanism in accordance with category learning models such as COVIS (Ashby et al., 1998) and ATRIUM (Erickson & Kruschke 1998). While the specifics of how multiple category systems interact can differ and it is not in the scope of this study to weight on this issue. Yet, it seems possible that some participants used a unidimensional rule while learning about both dimensions using an exemplar-based system.

Finally, it is noteworthy that while the number of boundary-based and typicality-based models considered in the DDM Analysis were equal (three difficulty maps each), forty participants were best fit by the boundary-based models while only three participants were best fit by the typicality-based models. This result is consistent with previous studies showing that two-alternative forced choice tasks with a verbalizable category distinction promotes boundary-based learning (Ell et al., 2017; Hélie et al., 2017). While contributing to the boundary-based vs typicality-based learning was not a goal of this study and different models were included for completeness, in general, the methods used in this article can be used to facilitate the identification of the type of knowledge that is gained under different training methodologies (e.g., A/B vs Yes/No training) for different category structures (e.g., rule-based vs information-integration).

### Methodological considerations

In the Learned Knowledge section of this article, it was assumed that a participant either learned a dimension or not. However, this assumption is not necessarily correct. For example it is possible for a participant to learn both the bar width and orientation features of “A” and “B” categories but only learn the bar width feature of “C” and “D” categories. In other words it is possible to classify participants in additional (more specific) clusters based on their learned knowledge. This more accurate clustering of participants was not included since using too many

clusters would require a larger sample of participants and the current clustering was sufficient to answer the research questions posed in the beginning of this article. Still, future work should be devoted to a larger-scale experiment that could include more specific clustering.

The main shortcoming of the iDBM Analysis was already mentioned in the Methods section, which is its restrictive assumption that participants were using unidimensional rules and switching between them. In other words, the model space of the iDBM Analysis was limited. However, we were careful in interpreting the results from this analysis, and the conclusions drawn from the iDBM Analysis were proportional to its limitations. The main conclusion from the iDBM Analysis is the existence of participants who learn features without relying on error feedback and therefore, similar to the Learned Knowledge section, we do not think that this shortcoming affects the conclusions.

Similar to the iDBM Analysis, one of the main shortcomings of the DDM Analysis is its model space. Even though the DDM Analysis considered a wider range of models compared to the iDBM Analysis, it is still possible to think of many more difficulty maps. The DDM Analysis however, takes this shortcoming into account to some degree by including the possibility of assigning a participant to none of the considered clusters (labeled as “None”). The “None” label covered possibilities that are not conceptually considered and the fact that forty-nine of the participants were assigned to it shows that there are indeed participants whose difficulty map are not similar to any of considered maps. The second issue with the DDM Analysis is the assumption that difficulty maps do not change in the last three blocks. This could explain why many participants did not fit well to any one of the difficulty maps. Overall, while each of the two analyses has shortcomings, the goal of this study was to advocate for using a single method to identify participants’ strategies, Instead, we argue for using multiple converging approaches to

show the opposing tendencies in category learning that can result in different learning outcomes for different individuals.

Finally, this study relied on a two-alternative forced choice task while in many real-life situations there are no explicit query to categorize objects and learning occurs in a supervised observational mode (Ashby et al., 2002). This is probably why most participants matched the boundary based models in the DDM Analysis and very few matched with the typicality based models as discussed in the previous section.

### Future work

A possible explanation for not learning both dimensions is limited attentional resources. If this is the case, will all participants learn both dimensions as the task becomes easier with practice? If yes, will all participants eventually use the same strategy? Or will there still be differences in the final strategy that participants settle on? Another question that can be asked is to investigate whether there is a link between the number of learned features under redundancy and cognitive characteristics of a participant. For example, participants that learned only one dimension may be better at inhibiting task irrelevant information. As a result, finding a stimulus dimension that works may have filtered the other dimension completely (so that differences between categories in the other dimension are never noticed). Future work is needed to determine which (if any) cognitive characteristics are predictive of the individual differences observed.

In addition to the non-methodological questions mentioned above, there are different aspects of the modeling that can be improved. There is nothing special about the difficulty maps included in this work other than that they were derived from popular categorization models. The proposed DDM Analysis can be used with other models or other types of data with minimal

changes. There is also no specification in this article on how the two dimensions are combined, which is something that can be incorporated by using the framework of Systems Factorial Technology (SFT) (Townsend & Nozawa, 1995). Another aspect of the DDM Analysis that can be improved is to allow variation in dimensional weighting. In this article, the attention is either completely on only one dimension (in the case of unidimensional difficulty maps) or is divided equally between the two dimensions. Allowing for variation in dimensional weighting would cover the attentional splits in between the two extremes considered here. With minor modifications, DDMs with trial-by-trial variability could be used as a general purpose tool in categorization. The concept of difficulty map can be used as a behavioral signature to test category learning models and previous experimental findings in new contexts. This would facilitate moving away from fitting average abstract participants (which may not exist) and instead learn about the individual participants that visited our labs. Finally, this study is relying on contrastive learning mechanisms (Davis & Love, 2010) to study redundancy by asking participants to distinguish only between specific pairs of categories in the training phase. More specifically, contrastive learning is used to study what dimensions are learned and used. However, the contrasts used in this study (or similar contrasts) can be used to answer more general questions about category representations and how using such contrasts affects them.

# References

Anderson, J. R. (1990). The adaptive character of thought.

Ashby, F. G. (2000). A stochastic version of general recognition theory. *Journal of Mathematical Psychology, 44*(2), 310-329.

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., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & cognition*, *30*(5), 666-677.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological review*, *95*(1), 124.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. Perception & Psychophysics, 61(6), 1178-1199.

Ashby, F. G., & Soto, F. A. (2015). Multidimensional signal detection theory. *Oxford handbook of computational and mathematical psychology*, 13-34.

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

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental psychology*, *36*(2), 190.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision*, *10*(4), 433-436.

Davis, T., & Love, B. C. (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science*, *21*(2), 234-242.

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. Nature neuroscience, 14(11), 1462-1467..

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, *64*(4), 1-34.

Ell, S. W., & Ashby, F. G. (2012). The impact of category separation on unsupervised categorization. Attention, Perception, & Psychophysics, 74(2), 466-475.

Ell, S. W., Smith, D. B., Deng, R., & Hélie, S. (2020). Learning and generalization of within-category representations in a rule-based category structure. *Attention, Perception, & Psychophysics*, *82*, 2448-2462.

Ell, S. W., Smith, D. B., Peralta, G., & Hélie, S. (2017). The impact of category structure and training methodology on learning and generalizing within-category representations. *Attention, Perception, & Psychophysics*, *79*(6), 1777-1794.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. Journal of Experimental Psychology: General, 127(2), 107.

Hélie, S. & Ashby, F. G. (2012). Learning and transfer of category knowledge in an indirect categorization task. *Psychological Research, 76,* 292-303.

Hélie, S., Turner, B. O., Crossley, M. J., Ell, S., & Ashby, F. G. (2017). Trial-by-trial identification of categorization strategy using iterative decision bound modeling. *Behavior Research Methods, 49,* 1146-1162.

Hélie, S., Shamloo, F., & Ell, S. W. (2017). The effect of training methodology on knowledge representation in categorization. *PloS one*, *12*(8), e0183904.

Hélie, S., Waldschmidt, J. G., & Ashby, F. G. (2010). Automaticity in rule-based and information-integration categorization. *Attention, Perception, & Psychophysics*, *72*(4), 1013-1031.

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

Maddox, W. T., Ashby, F. G., & Waldron, E. M. (2002). Multiple attention systems in perceptual categorization. *Memory & Cognition*, *30*(3), 325-339.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological bulletin*, *129*(4), 592.

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

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

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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, *85*(2), 59.

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision*, *2*(4), 237.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873-922.

Ratcliff, R., Perea, M., Colangelo, A., & Buchanan, L. (2004). A diffusion model account of normal and impaired readers. *Brain and Cognition*, *55*(2), 374-382.

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and aging*, *19*(2), 278.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive psychology*, *51*(1), 1-41.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, *64*(4), 583-639.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, *55*(4), 189.

Tolman, E. C., & Honzik, C. H. (1930). Introduction and removal of reward, and maze performance in rats. *University of California publications in psychology*.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology, 39*(4), 321-359.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics*, *7*, 14.

# Appendix A

# Simulations for DDM Analysis

In order to make the simulation useful in assessing the reliability of the result of this specific experiment, all the seven models were fit to all the participants and the group level estimates were used to generate random data. Table A1 shows the estimate for each of the models.

Table A1.

*The group level estimates of the DDM parameters.*

| | | Parameters | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | a | a_std | t | t_std | v | v_std | $\beta$ |
| Model | No Covariate | 2.23568 | 0.39393 | 0.34961 | 0.38255 | 1.27886 | 0.4431 | #N/A |
| | BW boundary | 2.23947 | 0.39497 | 0.34695 | 0.37987 | 1.28182 | 0.44475 | 0.08605 |
| | BW prototype | 2.23532 | 0.39439 | 0.34916 | 0.38198 | 1.27736 | 0.44427 | 0.03423 |
| | OR boundary | 2.2382 | 0.39444 | 0.34661 | 0.37872 | 1.28035 | 0.44722 | 0.07534 |
| | OR prototype | 2.2354 | 0.39383 | 0.34972 | 0.38281 | 1.27729 | 0.44456 | 0.02493 |
| | 2D boundary | 2.24258 | 0.39445 | 0.34963 | 0.38241 | 1.28622 | 0.4478 | 0.11037 |
| | 2D prototype | 2.23561 | 0.39298 | 0.34846 | 0.38075 | 1.27708 | 0.44455 | 0.0439 |

As Table A1 shows, the group level estimates for all the seven models were almost exactly the same with the exception of the estimates of $\beta$.The reason that $\beta$ estimates vary is that the individual differences that exist are manifesting themselves in $\beta$. For example, a participant

that best fits to the model that corresponds to the difficulty map of '2D boundary' model, is going to have a $\beta$ estimate closer to zero for all the other models which drags the group level $\beta$ estimate closer to zero. This means that a model with a difficulty map that corresponds to only a few participants will have a smaller group level $\beta$ estimate compared to a difficulty map that corresponds to more participants. As a result, we believe the individual level $\beta$ estimates corresponding to the correct map would be higher than the estimates in Table A1.

In order to test for this possibility, we fit the 'BW boundary' model to 'Learned_BW' participants and the 'OR boundary' model to 'Learned_OR' participants since they form a more homogenous group compared to the mix of all participants. The $\beta$ estimate of fitting 'BW boundary' model to 'Learned_BW' was 0.14 and the $\beta$ estimate of fitting 'OR boundary' model to 'Learned_OR' was 0.17.

Lastly, and as a result of the variability in $\beta$ estimates depending on how the model was fit, we decided to run three more simulations: One using $\beta = 0.06$ (mean of the $\beta$ estimates of all models fitted to all participants), another simulation using $\beta = 0.11$ (maximum of the $\beta$ estimates of all models fitted to all participants) and finally, a simulation using $\beta = 0.17$ (the $\beta$ estimate of fitting 'OR boundary' model to 'Learned_OR' participants). This simulation thus allows for quantifying the accuracy of the model selection process as a function of the magnitude of $\beta$. In each of the three simulations, the other parameters that were used to generate random data were the mean of the estimates shown in Table A1. The following procedure was used for each simulation: 288 (i.e., three blocks of 96 trials) random RT and response data was generated according to each difficulty map and the generated data was assigned to one of the labels using the procedure described in the 'Model selection process in DDM Analysis' section of the paper. This was repeated 100 times and therefore in the ends, there was 100 sets of data (each with 288

trials) for each of the seven models, for each value of $\beta$. The random data generated according to the 'No Covariate' model did not depend on $\beta$ and 97 out of 100 random data sets were correctly assigned to the 'No Covariate' model while three were assigned to 'None' label. The results of the rest of the simulated data according to the other six models is shown in Tables A2, A3, and A4. Each table is summarized using three numbers: number of correct labels, total number of 'vague' labels (i.e., assigned to 'No Covariate' and 'None'), and finally, number of misidentifications.

Simulation with the $\beta = 0.06$:

Table A2.

*The confusion table for the relation between identified model and the model that was used to generate the data with β=0.06.*

| | | Selected model | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | No Covariate | BW boundary | BW prototype | OR boundary | OR prototype | 2D boundary | 2D prototype | None |
| Simulate Data | No Covariate | 97 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| | BW boundary | 88 | 0 | 1 | 0 | 1 | 0 | 1 | 9 |
| | BW prototype | 90 | 0 | 3 | 0 | 0 | 0 | 0 | 7 |
| | OR boundary | 82 | 0 | 0 | 3 | 0 | 2 | 0 | 13 |
| | OR prototype | 79 | 0 | 0 | 0 | 4 | 1 | 1 | 15 |
| | 2D boundary | 83 | 1 | 1 | 1 | 0 | 1 | 1 | 12 |
| | 2D prototype | 87 | 0 | 2 | 0 | 0 | 0 | 0 | 11 |

Summary of the table (excluding data simulated according to the ‘No Covariate’ model):

Correct assignments: 11

Vague assignments: 576

Incorrect assignments: 13

Total number of simulations: 600

Simulation with the $\beta = 0.11$:
Table A3.

*The confusion table for the relation between identified model and the model that was used to generate the data with β=0.11.*

| | | Selected model | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | No Covariate | BW boundary | BW prototype | OR boundary | OR prototype | 2D boundary | 2D prototype | None |
| Simulate Data | No Covariate | 97 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| | BW boundary | 61 | 12 | 1 | 0 | 0 | 2 | 0 | 24 |
| | BW prototype | 58 | 0 | 17 | 0 | 0 | 0 | 3 | 22 |
| | OR boundary | 65 | 0 | 1 | 7 | 0 | 0 | 0 | 27 |
| | OR prototype | 73 | 0 | 0 | 1 | 8 | 0 | 1 | 17 |
| | 2D boundary | 63 | 2 | 0 | 2 | 0 | 1 | 0 | 32 |
| | 2D prototype | 60 | 0 | 0 | 1 | 4 | 0 | 8 | 27 |

Summary of the table (excluding data simulated according to the 'No Covariate' model):

Correct assignments: 53

Vague assignments: 529

Incorrect assignments: 18

Total number of simulations: 600

Simulation with the $\beta = 0.17$:

Table A4.

*The confusion table for the relation between identified model and the model that was used to generate the data with β=0.17.*

| | | Selected model | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | No Covariate | BW boundary | BW prototype | OR boundary | OR prototype | 2D boundary | 2D prototype | None |
| Simulate Data | No Covariate | 97 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| | BW boundary | 35 | 33 | 1 | 0 | 0 | 0 | 0 | 31 |
| | BW prototype | 34 | 0 | 32 | 0 | 0 | 0 | 3 | 31 |
| | OR boundary | 41 | 0 | 0 | 25 | 0 | 1 | 0 | 33 |
| | OR prototype | 31 | 0 | 0 | 0 | 38 | 0 | 1 | 30 |
| | 2D boundary | 33 | 2 | 0 | 2 | 0 | 17 | 0 | 46 |
| | 2D prototype | 28 | 0 | 2 | 0 | 1 | 0 | 22 | 47 |

Summary of the table (excluding data simulated according to the 'No Covariate' model):

Correct assignments: 167

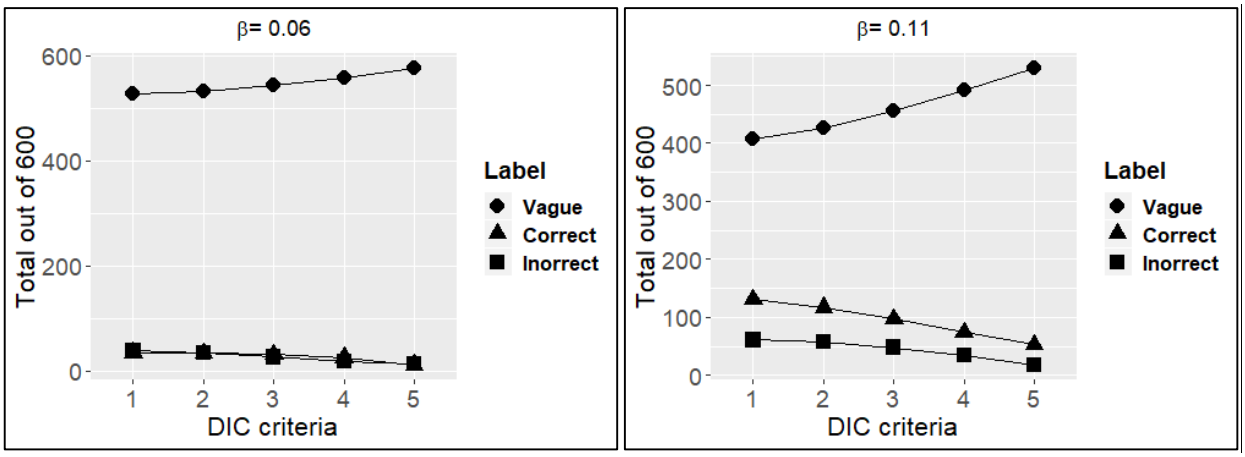
Vague assignments: 420

Incorrect assignments: 13

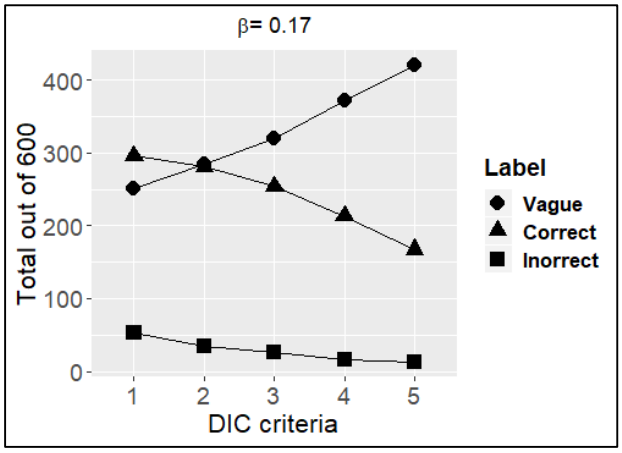
Total number of simulations: 600

The above simulations were ran using different threshold for DIC scores but only the threshold five (used in this article) is shown in the tables above. The specific threshold was chosen to minimize the number of incorrect identifications. Figures A1a, A1b and A1c visually

display the results under different DIC thresholds for $\beta = 0.06$, $\beta = 0.11$ and $\beta = 0.17$ respectively.




(a) (b)




(c)

*Figure A1*. The relation between DIC threshold for model selection and number of correct, incorrect and vague model identifications.

Figure A1 shows that if the signal is weak ($\beta = 0.06$), then it is not possible to detect any meaningful patterns and almost all participants are assigned to the Vague labels. However, with

the stronger signal ($\beta = 0.17$), there is a tradeoff between Vague and Incorrect assignments. The higher the threshold for DIC, the more vague assignments but less Incorrect assignments. A conservative DIC threshold of five was chosen since the goal of this article is to show the existence of different clusters of participants (e.g., showing that a participant that is learning both dimensions can be using only one of them). For this reason, it is important to be confident that an incorrect identification is avoided. Therefore, even though choosing five as the threshold results in many vague identifications, it was chosen as the threshold.