Supervised category learning: When do participants use a partially diagnostic feature?

Sujith Thomas (sujitht@goa.bits-pilani.ac.in)
Department of Computer Science & Information Systems, BITS Pilani, K. K. Birla Goa Campus, India 403726

Aditya Kapoor (f20170963@goa.bits-pilani.ac.in)
Department of Computer Science & Information Systems, BITS Pilani, K. K. Birla Goa Campus, India 403726

Narayanan Srinivasan (nsrini@iitk.ac.in)
Department of Cognitive Science, Indian Institute of Technology Kanpur, India 208016

Abstract

We report a supervised category learning experiment in which the training phase contains both classification and observation learning blocks. To explain the use of different categorization strategies, we propose an account in which use of a stimuli dimension depends on how well the dimension is learned. Our results show that there is an overall preference for a unidimensional categorization based on the perfectly diagnostic dimension. The preference for unidimensional categorization is negatively correlated with how well participants learn the partially diagnostic dimensions. Preference for unidimensional categorization is also negatively correlated with the mean response time. Bayesian modeling results show that participants use a partially diagnostic dimension only when it is learned with a very high level of accuracy. Different strategies are used for categorization depending on how well the perfectly and partially diagnostic dimensions are learned.

Keywords: supervised category learning; classification learning; observation learning; Bayesian modeling

Introduction

Classification learning and Observation learning are two types supervised category learning paradigms commonly used in the literature (Nelson, 1984; Levering & Kurtz, 2015). In classification learning, instances of two different categories are presented one by one and participants are asked to classify each instance. Feedback is given and participants learn through trial and error. In observation learning, instance of a category is presented along with its category label. Participants observe each labelled instance before moving to the next instance. No feedback is given in observation learning.

Studies have consistently reported that in classification learning participants preferred a unidimensional categorization based on the perfectly diagnostic feature (Nelson, 1984; Rabi, Miles, & Minda, 2015). For observation learning, the results showed that there was a greater preference for a similarity based strategy that was based on all the diagnostic features (family resemblance structure) of a category (Nelson, 1984; Smith & Shapiro, 1989). However, Ward and Scott (1987) showed that participants preferred a unidimensional rule plus exception strategy for both classification and observation learning.

Levering and Kurtz (2015) showed that the perfectly diagnostic dimension was learned with 100% accuracy for both observation and classification learning. Levering and Kurtz (2015) also showed that the accuracy for the partially diagnostic features was significantly higher in observation learning (85%) compared to classification learning (72%). We think that a greater preference for a similarity based strategy in some observation learning studies is due to better learning of the partially diagnostic features. For example, Thomas and Srinivasan (2020) showed that the preference for a unidimensional strategy (based on the perfectly diagnostic feature) decreased when participants were made to memorize the partially diagnostic features.

This study aims to check whether there is a relation between preference for a unidimensional strategy and how well partially diagnostic features are learned. In our study, we have used both observation and classification learning in the training phase. We show that the perfectly diagnostic dimension is learned better compared to the partially diagnostic dimensions, which is consistent with the results by Levering and Kurtz (2015). We show that the percentage of unidimensional categorization is negatively correlated with how well the partially diagnostic features are learned; also, the participants whose average response time was longer made fewer unidimensional responses. We use Bayesian modeling to show that participants use a dimension only when they have learned the dimension with a high level of accuracy. We use the results of Bayesian modeling to propose an alternative theoretical position where the preference for unidimensional strategy is stronger when accurate knowledge about diagnosticities of features are not available. We argue that this alternative theoretical position has a broader explanatory power.

In the rest of this article, we refer to the categorization strategy based on the perfectly diagnostic feature as the CA (criterion attribute) strategy, and the categorization strategy based on all the diagnostic features as the FR (family resemblance) strategy. We refer to the perfectly diagnostic dimension as CA (criterion attribute) dimension, and we refer to the partially diagnostic dimensions as FR (family resemblance) dimensions.

Experiment

Thomas and Srinivasan (2020) reported that participants showed a preference for the CA strategy even when both observation and classification learning blocks were used in the training phase. In this experiment, we wanted to replicate the results when there are no explicit memorization conditions. We hypothesized that there would be a correlation between preference for the CA strategy and how well the FR features are learned.

Method

Subjects Forty five volunteers (5 females; mean age = 21.5 years) participated in this experiment. The results of linear regression reported by Thomas and Srinivasan (2020) indicated that there was an effect between the accuracy for FR features and percentage of CA categorization ($R^2 = .16$ in condition M0). Power analysis (Bausell & Li, 2002) indicated 44 subjects are needed for power = .80, R^2 = .16 and two-tailed α =.05 for a significant effect between the accuracy for FR features and percentage of CA categorization.

Materials Figure 1 shows the fish-like stimuli that were used in Experiment 1. The stimuli consisted of five dimensions — shape of the tail, shape of the upper-fin, shape of the lower-fin, shape of the mouth and the body pattern. Each stimuli dimension could take one of two possible values. One of the five feature dimensions was perfectly diagnostic of category membership (CA dimension); the remaining four feature dimensions were partially diagnostic (FR dimensions). In Figure 1 shape of the mouth is the CA dimension and shape of the tail is one of the four FR dimensions. The stimuli in the first two rows of Figure 1 formed the training stimuli. The stimuli in the last two rows formed the transfer stimuli. The transfer stimuli were constructed by flipping the CA feature of the training stimuli. In other words, each transfer stimulus contained the CA feature and FR features from opposite categories.

We used five different sets of stimuli, where a different stimuli dimension formed the CA dimension in each set. For each participant, one of the five sets of stimuli was used. Each of the five stimuli sets was used 9 times in the experiment (9×5 sets $= 45$ participants). In each of the five sets, the CA dimension was always black in colour, two FR dimensions were always yellow, and the remaining two were always blue. The colour on its own did not help in identifying the categories (as can be see in Figure 1). Colours can make the FR dimensions more salient, but a pilot study revealed that participants continued to show a strong preference for the CA strategy. In the extended version of this study, we plan to run an experiment where colours covary with the FR features.

Procedure We developed a web application using the Django framework for collecting the behavioral data. The link to the web application was sent to participants over an email. Participants responded in a self-paced manner throughout the experiment.

The experiment started with a training phase. In the training phase, participants were asked to learn to differentiate category A objects from category B objects. The instructions given to the participants were neutral and did not indicate the categorization strategy that they were expected to use. Each block in the training phase consisted of an observation learning sub-block followed by a feedback learning sub-block. In the observation learning sub-block, the 10 training stimuli were presented one by one along with the correct category label. In the classification learning sub-block, the 10 training stimuli were again presented one by one and participants were asked to categorize each stimulus. Feedback was given after every response and indicated whether the response was correct. At the end of the classification learning sub-block, participants were shown their accuracy for the classification learning sub-block. Participants had to achieve an accuracy of 90% twice (learning criterion) in order to proceed to the next phase. The training phase was repeated until participants could achieve an accuracy of 90% two times.

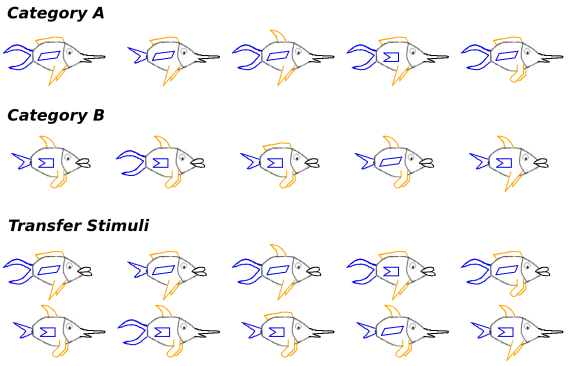

Figure 1: Training stimuli and transfer stimuli used in the experiment. Here, the shape of the mouth is the perfectly diagnostic (CA) dimension, and the remaining features are partially diagnostic (FR) features.

The training phase was followed by the transfer phase. In each block of the transfer phase, all the 10 transfer stimuli were presented one by one and participants were asked to categorize each stimulus. No feedback was given. The transfer phase contained three blocks. So, participants categorized each transfer stimulus three times.

The transfer phase was followed by an all features test phase, where participants were asked to identify the category in which a given feature occurred more commonly in. Participants were not informed that there will be an all features test phase. There were two features along each of the five stimuli dimensions. So, there were ten features in total. In one block, all the ten features were tested once. No feedback was given. There were three blocks in the all features test phase.

After the all features test phase, participants were asked to describe how they categorized the items.They were requested to describe their strategy with sufficient clarity so that another person may read the description and replicate their categorization pattern.

Results

In the transfer phase, 30 participants (out of 45) preferred the CA strategy more than 90% of the times. A one-sample t-test showed that the overall percentage of CA categorization ($M = 74.44\%, SD = 38.03$) was above the chance level (50%); $t(44) = 4.26, p = 0.0001, d = 0.64$. This shows that there was a strong preference for the CA strategy.

In the all features test phase, the mean and standard devia-

Table 1: The four columns represent the groups based on participants' categorization descriptions. The rows show the number of participants in each group (row 1), the average CA categorization responses in the transfer phase (row 2), mean response time for the transfer stimuli (row 3), mean accuracy for the CA dimension in the all features test phase (row 4), mean accuracy for the FR dimensions (row 5) and the average number of FR dimensions for which a participant achieved 100% accuracy in the all features test phase (row 6).

	CA	MULTI	FR	OTHER
1. No. of participants	28	7	6	4
2. CA strategy	97%	48%	29%	36%
3. Mean RT	1.6s	4.5s	9.1s	3.0s
4. CA accuracy	96%	90%	95%	67%
5. FR accuracy	73%	77%	87%	80%
6. FR 100% accuracy	1.6	2.0	3.0	2.0

tion for the CA and FR dimensions were $(M = 92.59\%, SD = 19.10)$ and $(M = 76.02\%, SD = 17.35)$ respectively. A one-sample t-test showed that the accuracy for the CA dimension was above chance level (50%); $t(44) = 14.79, p < .0001, d = 2.20$. The accuracy for the FR dimension was also above chance level (50%); $t(44) = 9.95, p < .0001, d = 1.48$. A paired-sample t-test showed that the difference in accuracy for the CA and FR dimensions was significant; $t(44) = 4.47, p < .0001, d = .67$. This shows that participants learned the CA dimension better than the FR dimensions.

To further understand feature learning and categorization strategy, participants were divided into four groups based on their categorization descriptions (Table 1). Participants who used only the CA dimension were labeled as CA (column 2 in Table 1). Participants who used two to four stimuli dimensions were labeled as MULTI (column 3). Participants who used all the five dimensions were labeled as FR (column 4). Participants who used other strategies (like counting number of pointy features) were labeled as OTHER (column 5). Some participants used the CA strategy 100% of the time, but described their strategy to be multidimensional. We put such participants in the MULTI group (and not in CA). Due to this, the number of participants in the CA group is 28, and not 30 as reported above. Importantly, no participant claimed to have used a unidimensional strategy based on a single FR dimension. This is possibly due to the learning criterion of 90% accuracy in the training phase, which cannot be achieved using a unidimensional strategy based on a single FR dimension.

The third row shows that the mean RT for the transfer stimuli was more for the MULTI and FR groups compared to the CA group. The sixth row in Table 1 shows the average number of FR dimensions with 100% accuracy (i.e. no error). On average, participants in the FR group could learn 3 FR dimensions with 100% accuracy, whereas this number was only 1.6 for the CA group. This data becomes relevant in the light of our results in the Bayesian modeling section.

Table 1 suggests that there is a correlation between CA strategy and accuracy for the FR dimensions. The results of linear regression indicated that the accuracy for the FR features is a significant predictor of the percentage of CA categorization, $\beta = -.32$, $t(43) = 2.24, p = .03, R^2 = .11$, adjusted $R^2 = .08$. The percentage of CA strategy decreased with increase in accuracy for the FR dimensions.

Table 1 also suggests that there is a correlation between the CA strategy and the mean RT. The results of linear regression indicated that the mean RT is a significant predictor of CA categorization percentage, $\beta = -.61, t(43) = 5.11, p < .0001$, $R^2 = .38$, adjusted $R^2 = .36$. The percentage of CA strategy decreased with increase in the mean RT. In fact, none of the participants whose mean RT was more than 4s showed a preference for the CA strategy. These results are consistent with the finding that a multidimensional strategy is more effortful and takes more time (Milton, Longmore, & Wills, 2008; Wills, Milton, Longmore, Hester, & Robinson, 2013).

Bayesian Modeling

In classification learning, it has been shown that participants show a strong preference for the CA strategy (Nelson, 1984; Rabi et al., 2015). This has been explained as follows:
Theoretical position A. Classification learning promotes analytical processing of information. Due to this, participants test unidimensional rules until they find the best rule (Nelson, 1984). This leads to a greater preference for the CA strategy and FR features are not learned well.

Our results show that 74.44% of the overall categorization responses were based on the CA strategy. This can be explained using theoretical position A. However, we also found that participants with a higher accuracy for the FR dimensions showed less preference for the CA strategy. We used Bayesian modeling to check whether there is an accuracy threshold below which participants ignore the FR dimensions.

Figure 2 shows the Bayesian model that we have used. The shaded nodes represent the observed variables. The unshaded nodes having a single border represent the free parameters whose value depends on a prior probability distribution. The square-shaped nodes take discrete values, while circular nodes take continuous values. The nodes with double borders are the deterministic nodes whose value depend on the parent node(s).

We have used three observed variables: $\vec{x_i}$, $\vec{t_k}$ and r^A_{ki}. The variable $\vec{x_i}$ is a five dimensional vector that corresponds to the *logical* representation of the i^{th} transfer stimulus. As described earlier, we have used five different sets of stimuli, where each set had a different stimuli dimension forming the CA dimension. We have used a logical representation in which the first dimension is always the CA dimension. Suppose that for a stimuli set the third stimuli dimension is the CA dimension. Then the third stimuli dimension is mapped to the first logical dimension, the fourth stimuli dimension is mapped to the second logical dimension and so on. After

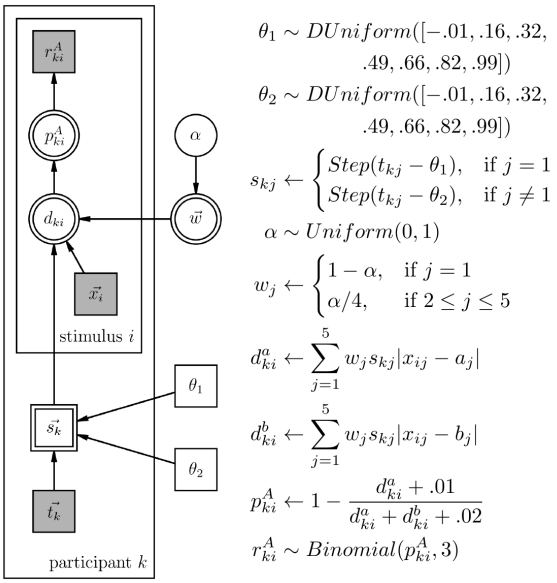

Figure 2: Graphical representation describes the stochastic processes that generate the observed data from unobserved parameters for model I.

the fifth stimuli dimension we rotate back to the first stimuli dimension and continue the mapping.

The second observed variable that we have used is $\vec{t}_k$, which is also a five dimensional vector. The j^{th} component of vector $\vec{t}_k$ (denoted by t_{kj}) contains the accuracy of the k^{th} participant for the j^{th} *logical* dimension. This means that the first component of vector $\vec{t}_k$ will always contain the accuracy for the CA dimension. In each block of the all features test phase, every stimuli dimension was tested twice (because each dimension has two features). In total there were three blocks. This means that each stimuli dimension was tested six times in the all features test phase. The accuracy for the j^{th} *logical* dimension for the k^{th} participant (i.e. t_{kj}) was found by dividing the number of correct responses for each dimension by 6. So, t_{kj} can have one of the following discrete values: 0, .17, .33, .5, .67, .83 or 1. The accuracy values have been rounded to two decimal places.

The third observed variable r^A_{ki} tells us how many times the k^{th} participant categorized the i^{th} transfer stimulus to category A. In the transfer phase, there were three blocks. So, every participant categorized every transfer stimulus three times. For this reason, the value of variable r^A_{ki} will be 0, 1, 2 or 3. The deterministic parameter p^A_{ki} gives the probability that k^{th} participant will categorize a transfer stimulus $\vec{x}_i$ to category A. In our model, the observed variable r^A_{ki} is generated from a Binomial distribution with probability p^A_{ki} and three Bernoulli trials (see Figure 2).

The probability p^A_{ki} was determined using distances d^a_{ki} and d^b_{ki}. Distance d^a_{ki} is the distance of stimulus $\vec{x}_i$ from the prototype of category A (denoted by $\vec{a}$). The prototype $\vec{a}$ will contain the diagnostic features (both CA and FR) of category A. So the distance d^a_{ki} will increase if stimulus $\vec{x}_i$ contains fewer diagnostic features of category A. We have found the distance d^A_{ki} as follows:

$$d^a_{ki} = \sum_{j=1}^{5} w_j s_{kj} |x_{ij} - a_j| \tag{1}$$

where w_j is the attention weight for logical dimension j and s_{kj} is a parameter that can be either 0 or 1. The parameters w_j and s_{kj} are explained below. In Eqn. (1), $|x_{ij} - a_j|$ will be 0 if $\vec{x}_i$ and $\vec{a}$ have the same j^{th} feature, otherwise $|x_{ij} - a_j|$ will be 1. We find the distance d^b_{ki} in a manner similar to Eqn. (1), where $\vec{b}$ denotes the prototype of category B.

In Eqn. 1, parameter s_{kj} determines whether the k^{th} participant used the j^{th} dimension for categorization. We have hypothesized that participant k would use the j^{th} dimension only when the accuracy t_{kj} is above some threshold (θ). To capture this relation we have used a unit $Step()$ function.

$$s_{kj} \leftarrow \begin{cases} Step(t_{kj} - \theta_1), & \text{if } j = 1 \\ Step(t_{kj} - \theta_2), & \text{if } 2 \leq j \leq 5 \end{cases} \tag{2}$$

The value of the $Step()$ function is 1 when its argument is positive, otherwise it is 0. So, parameter s_{kj} will be 1 if the accuracy t_{kj} is above some threshold, otherwise s_{kj} will be 0. If s_{kj} is 0, then the j^{th} logical dimension will be ignored while computing the distance in Eqn. (1). In Eqn (2), we have used threshold θ_1 for the CA dimension (i.e. $j = 1$), and threshold θ_2 for the four FR dimensions (i.e. $2 \leq j \leq 5$). This is because we are working with the logical representation where different stimuli dimensions can map to the same logical dimension depending on the stimuli set being used. For this reason, we did not want to differentiate between the different FR dimensions and have used the same threshold θ_2.

The threshold parameters (θ_1 and θ_2) are generated from a uniform prior distribution over the following discrete values: $-.01$, .16, .32, .49, .66, .82 and .99. These values are .01 less than the accuracy values that t_{kj} can take. We have used discrete values for the threshold because the accuracy values are also discrete [1] .

In Eqn. (1), we have used attentional weights w_j for each dimension. These weights are determined as follows:

$$w_j \leftarrow \begin{cases} 1 - \alpha, & \text{if } j = 1 \\ \alpha/4, & \text{if } 2 \leq j \leq 5 \end{cases} \tag{3}$$

where α is a parameter having a uniform prior distribution in the interval $(0, 1)$. Determining the weights in the above manner ensures that all the weights sum to 1. Eqn. (3) allows the weights for the CA dimension to be different from the FR

[1] We could have used a continuous prior distribution (i.e $Uniform(0, 1)$) for the threshold parameters. In that case, the posterior distribution would lie in the open interval $(\frac{4}{6}, \frac{5}{6})$ and $(\frac{5}{6}, 1)$ instead of being concentrated at .82 and .99 respectively (See Figure 3). Since we are only interested in the minimum accuracy with which a dimension should be learned, we have used discrete values.

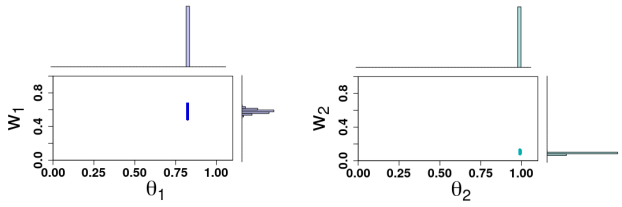

Figure 3: Posterior distribution for CA attentional weight (w_1), CA accuracy threshold (θ_1), FR attentional weight (w_2) and FR accuracy threshold (θ_2) for model I shown in Figure 2.

dimensions. However, the attentional weights for the FR dimensions are the same. The reason for this is (again) that we don't want to differentiate between different FR dimensions in the logical representation for the stimuli.

If the distance d^a_{ki} is small, it would mean that stimulus $\vec{x}_i$ has many diagnostic features of category A; therefore, the probability p^A_{ki} must be closer to 1. The probability p^A_{ki} was determined from distances d^a_{ki} and d^b_{ki} as follows:

$$p^A_{ki} = 1 - \frac{d^a_{ki} + .01}{d^a_{ki} + d^b_{ki} + .02} \tag{4}$$

where d^a_{ki} and d^b_{ki} lie in the range 0 to 1. The distances d^a_{ki} and d^b_{ki} can both become zero for a participant who has learned all the dimensions poorly (low accuracy), because s_{kj} will be 0 for all the dimensions. To avoid the divide-by-zero error in Eqn. (4), we have added constant values .01 and .02 in the numerator and the denominator respectively [2]. These values are small compared to the distances and will not have much effect when the distances are non-zero. If both the distances become zero, then the probability p^A_i will become .5 because of the constant values. This will model the fact that the participant who is not sure of the diagnosticity of any of the dimensions is probably giving random categorization responses.

Results

We have used the R2jags library (Su & Yajima, 2012) in R to obtain samples from the joint posterior distribution of parameters using Gibbs sampling. We have monitored the attentional weights ($\vec{w}$) and the accuracy thresholds (θ_1 and θ_2). All our results are based on three chains of 4,000 samples each (total 12,000 samples). Each of the three chains had a burn-in of 1,000 samples, and the samples were thinned by taking every tenth sample. The convergence of the three chains were checked using the standard $\hat{R}$ statistic (Brooks & Gelman, 1998).

Our model has three free parameters: α, θ_1 and θ_2. We had 450 data points (45 participants $\times$ 10 transfer stimuli). Figure 3 shows the posterior distribution for the parameters w_1, θ_1, w_2 and θ_2. The marginal posterior distribution for the threshold parameters θ_1 and θ_2 was concentrated at .82 ($M = .82, SD = 0$) and at .99 ($M = .99, SD = 0$) respectively. These results indicate that when participants use the j^{th} dimension, they are sure of the diagnosticity of the features along that dimension (i.e. $t_{kj} > .82$). The means and standard deviation for parameters w_1 and w_2 were ($M = .58, SD = .02$) and ($M = .10, SD = .01$) respectively. The attentional weight for the CA dimension (w_1) was greater than the attentional weight for the FR dimensions (w_2), which shows participants paid more attention to the CA dimension.

We compared two theoretical positions T_1 and T_0. Theoretical position T_1 states that an FR dimension will be used only when it is learned with a high level of accuracy (i.e. $t_{kj} > .99$). We modeled this theoretical position by letting the prior distribution for θ_2 to be concentrated at .99 (i.e. $\theta_2 \sim DUniform([.99])$). Theoretical position T_0 states that an FR dimension may be used even when it is not learned with a high level of accuracy. Theoretical position T_0 is the null hypothesis for T_1 and was modeled by letting the prior distribution for θ_2 to be $DUniform([-.01, .16, .32, .49, .66, .82])$. Note that in model T_0 we don't allow θ_2 to take a value of .99. All the other details of models T_1 and T_0 were same as model I shown in Figure 2. To reiterate, Model T_0 allows an FR dimension to be used at *any* level of accuracy, whereas for model T_1 the accuracy must be high ($t_{kj} > .99$).

We used Bayes factor (Kass & Raftery, 1995) to compare model T_1 and model T_0. Bayes factor was found as follows:

$$BF_{10} = \frac{P(Data|model\ T_1)}{P(Data|model\ T_0)} \tag{5}$$

where $P(Data|model\ T_i)$ is the marginal probability of generating the data given model T_i. The marginal probability was calculated by averaging over the 12,000 samples generated from the joint posterior distribution of the parameters using the same procedure described for model I.

The results of Bayes factor comparison between model T_1 and model T_0 shows that model T_1 provides a better explanation for the data ($BF_{10} > 100$, extreme evidence). This means that the participants who used an FR dimension were highly accurate for that dimension.

Next we checked whether the data can be explained purely based on theoretical position A described earlier. If we strictly follow the theoretical position A, then participants must always use the CA dimension because that is the best analytical solution to the classification problem. Also, the best analytical answer would remain the same irrespective of how well participants learn the FR dimensions. For this reason, we have modelled theoretical position A by setting the attentional weights for the FR dimensions in model T_1 to zero. So, in model T_A parameter α is always set to 0. All the other details of model T_A are the same as that of model T_1. We have modified model T_1 to obtain model T_A because we wanted to compare two models that differ in just one parameter.

The results of Bayes factor comparison between model T_1 and model T_A shows that model T_1 provided a better explanation for the data ($BF_{1A} > 100$, extreme evidence). This shows

[2] Instead of .01 and .02 we have tried using c and $2 \times c$, where c is a free parameter. The results obtained were identical. By using constants .01 and .02 we have tried to minimize the free parameters.

that the data cannot be explained purely on the basis of analytical processing. Participants used the FR dimensions, but only when they were highly accurate for those dimensions.

The conclusions drawn from the results of Bayesian modeling are consistent with the data shown in Table 1. The average number of FR dimensions learned with 100% accuracy (sixth row in Table 1) is lowest for the CA group and is highest for the FR group. In the groups shown in Table 1, the percentage of CA responses decreases as more FR dimensions are learned with 100% accuracy.

Discussion and Conclusion

Our results show that participants use a dimension only when it is learned with a high level of accuracy. Participants who deviated the most from the CA strategy had a high accuracy for several FR dimensions (see Table 1). Rabi et al. (2015) have reported similar results for classification learning, where participants showed an overall preference for the CA strategy. However, four participants who performed the best in the all features test preferred the FR strategy (Rabi et al., 2015, p.164). Our results are also consistent with the finding that a multidimensional strategy is more effortful and takes more time compared to a unidimensional strategy (Milton et al., 2008; Wills et al., 2013).

We have used distance from the category prototypes to predict the probability of categorizing a stimulus to category A. However, we do not argue in favor of prototype-based theories. The set of stimuli we have used does not help us differentiate between the prototype and exemplar theories. We only make a claim that some stimuli dimensions will be ignored during categorization, unless their diagnosticity is learned with a high level of accuracy.

Our results have non-trivial implications. Consider two groups of participants (X and Y) having three participants each. Let participants in group X have an accuracy of 60%, 65% and 100% for an FR dimension. Let group Y participants have an accuracy of 53%, 85% and 85% for the same FR dimension. The mean and SD of the two groups are similar. However, one participant in group X (having 100% accuracy) is much more likely to use the FR dimension compared to the other participants. This pattern in the data gets hidden when we look only at the mean and SD values of accuracy. Our results indicate the importance of checking whether individual participants have learned a dimension with a high level of accuracy.

A similar effect happens when a computational model has attentional weight parameters. During model fitting, these weights get adjusted depending on how many participants have used a stimuli dimension on an average. Once again the underlying pattern, in which some participants completely ignore a stimuli dimension, gets hidden.

Here, we propose an alternative theory (*Theoretical position B*), which postulates that participants will use the FR dimensions only when they have learned it with high accuracy and if not, will use a unidimensional strategy because it is less effortful. Theoretical position B can explain the preference for CA strategy in classification learning. Initially, participants do not know the diagnosticity of any feature. So, they use the less effortful unidimensional strategy. Theoretical position B can also predict the results of observational learning. Observational learning can lead to better learning of the FR dimensions (Levering & Kurtz, 2015). Model T_1 predicts that when more participants accurately learn the FR dimensions, there will be a lesser preference for the CA strategy. This can explain why some observational studies report a preference for a unidimensional categorization (Ward & Scott, 1987).

In *match-to-standards* procedure, participants are shown the prototypes of two categories and told that the prototypes belong to opposite categories (Regehr & Brooks, 1995; Milton et al., 2008). Participants are then asked to categorize a transfer stimulus [3]. In this experimental procedure, accurate knowledge of the diagnosticities of all the dimensions are (perceptually) available to a participant. Theoretical position B predicts that in such a situation there would be a lesser preference for unidimensional categorization and a greater preference for FR categorization. The results for the *match-to-standards* procedure (Regehr & Brooks, 1995; Milton et al., 2008) are consistent with this prediction.

Accuracy of stimuli dimensions are contingent on the attention allocated to the stimuli dimensions. Rehder and Hoffman (2005) have shown that in classification learning participants tend to allocate attention to stimuli dimensions in a manner that optimizes category discrimination. Once the errors get sufficiently reduced, participants start allocating more attention to other stimuli dimensions as well. The learning criterion of achieving 90% accuracy twice could have enabled participants to allocate more attention to the partially diagnostic features, thereby learning those features better. The results reported by Rehder and Hoffman (2005) predicts that a more lenient learning criterion might lead to a greater preference for the CA strategy.

The attention allocated to the partially diagnostic dimensions also depends on the experimental procedure used. In the training phase of inference learning (Yamauchi & Markman, 1998; Chin-Parker & Ross, 2004), each stimulus is presented along with it category label, but one of the features is missing. Participants need to correctly predict the missing feature. Feedback is provided. The results showed that inference learning led to better accuracy for the partially diagnostic features compared to classification learning. Also, inference learners rated the typicality of an item based on both the perfectly and partially diagnostic features, but the classification learners relied solely on the perfectly diagnostic features (Chin-Parker & Ross, 2004). These results are consistent with our finding that participants use a partially diagnostic dimension only when it is learned more accurately. Fur-

[3] *Match-to-standards* procedure is often considered to be an example of unsupervised categorization. However, it can also be thought of as a special case of supervised categorization where participants must generalize after being shown one member of each category.

ther studies need to determine whether the pattern revealed by Model T_1 also exists for classification and inference learning.

The categorization descriptions given by participants suggest that participants have ignored the colour information. Using colours for partially diagnostic features should make it more salient. However, our results are similar to the results reported for the black version of the same stimuli (Thomas & Srinivasan, 2020). In the extended version of this paper, we will be reporting the results of an experiment where colours covary with the partially diagnostic features.

It has been shown that for some types of two dimensional stimuli, similarity-based grouping takes less time (Ward, 1983). Also, salience of features have an effect on categorization strategy (Hammer, Sloutsky, & Grill-Spector, 2012). The nature of instructions given during the experimental procedure also influences the stimuli dimensions used for categorization (Medin & Smith, 1981; Kurtz, Levering, Stanton, Romero, & Morris, 2013). Theoretical position B cannot explain the effect due to these (and possibly other) factors. However, we believe that theoretical position B has better explanatory power compared to theoretical position A.

Thomas and Srinivasan (2020) have used an explicit manipulation where participants were made to memorize the FR features of a category. The results showed that as participants learned the FR features better there was a corresponding decrease in the preference for the CA strategy. This result is also consistent with theoretical position B. To conclude, our current empirical and modeling results show the flexible use of categorization strategies depending on the dimensions that were learned accurately.

References

Bausell, R. B., & Li, Y.-F. (2002). *Power analysis for experimental research: a practical guide for the biological, medical and social sciences*. Cambridge University Press.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, *7*(4), 434–455.

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: a comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 216.

Hammer, R., Sloutsky, V., & Grill-Spector, K. (2012). The interplay between feature-saliency and feedback information in visual category learning tasks. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 54–59). Austin, TX: Cognitive Science Society.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of shepard, hovland, and jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 552.

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & cognition*, *43*(2), 266–282.

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(4), 241.

Milton, F., Longmore, C. A., & Wills, A. (2008). Processes of overall similarity sorting in free classification. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 676.

Nelson, D. G. K. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior*, *23*(6), 734–759.

Rabi, R., Miles, S. J., & Minda, J. P. (2015). Learning categories via rules and similarity: Comparing adults and children. *Journal of experimental child psychology*, *131*, 149–169.

Regehr, G., & Brooks, L. R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 347.

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

Smith, J. D., & Shapiro, J. H. (1989). The occurrence of holistic categorization. *Journal of Memory and Language*, *28*(4), 386–399.

Su, Y.-S., & Yajima, M. (2012). R2jags: A package for running jags from r. *R package version 0.03-08, URL http://CRAN. R-project. org/package= R2jags*.

Thomas, S., & Srinivasan, N. (2020). Better learning of partially diagnostic features leads to less unidimensional categorization in supervised category learning. In Y. X. S. Denison. M. Mack & B. Armstrong (Eds.), *Proceedings of the 42nd annual conference of the cognitive science society* (pp. 3444–3450). Cognitive Science Society.

Ward, T. B. (1983). Response tempo and separable-integral responding: Evidence for an integral-to-separable processing sequence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(1), 103.

Ward, T. B., & Scott, J. (1987). Analytic and holistic modes of learning family-resemblance concepts. *Memory & Cognition*, *15*(1), 42–54.

Wills, A. J., Milton, F., Longmore, C. A., Hester, S., & Robinson, J. (2013). Is overall similarity classification less effortful than single-dimension classification? *The Quarterly Journal of Experimental Psychology*, *66*(2), 299–318.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and language*, *39*(1), 124–148.