Conditions conducive to optimal probabilistic categorization

John F. Ackermann[a], **Michelle A. Borkin**[b], **and Peter J. Bex**[a]

[a] Department of Psychology, Northeastern University, Boston, MA; [b] College of Computer and Information Science, Northeastern University, Boston, MA

This manuscript was compiled on August 11, 2017

In a probabilistic categorization task, each stimulus can occur in each category with some probability. This type of task is common and often of critical importance as in, for example, medical diagnosis. If the goal is to maximize accuracy, optimal categorization is achieved by choosing the highest probability (i.e., the maximum *a posteriori* (MAP)) category for a stimulus every time it is encountered. In this study we investigate the conditions that make an observer likely to adopt an optimal categorization strategy. Subjects were presented with stimuli defined by 3 discrete, binary features that were members of two possible categories. They first completed a Training block in which they were presented with a set of stimuli and their true category labels. Then, in a Testing Block, they categorized unlabeled stimuli. We compared human subjects' performance to that of two models. The results are well accounted for by a model that attributes optimality to the observer's uncertainty with regard to the optimal decision rule. Increases in uncertainty are dependent on 3 factors: 1) A decrease of the magnitude of the stimulus' MAP probability, 2) A decrease in the frequency with which the stimulus has been encountered in training and, 3) an increase in the observer's uncertainty with regard to the category-relevant stimulus dimensions. We also show that a measure of uncertainty involving stimulus probability, frequency, and similarity measured in its objective feature space can account for the relative degree to which subjects will adopt the optimal strategy.

Categorization | Optimality | Decision making

A probabilistic categorization task is one in which stimuli must be categorized on the basis of features that are not deterministically associated with the categories. That is, each stimulus can occur in each category with some probability. Medical diagnosis is an example [1]. In a diagnostic situation, the stimulus to be categorized is the set of observed symptoms presented by the patient or the outcome of a clinical test. A doctor attempts to categorize the patient as "sick" or "well" based on their observations. A given symptom, a headache, for example, can be indicative of serious illness or of no particular concern and a positive clinical test can be a false indication of disease. Thus the observed symptoms have only a probabilistic, and not a deterministic, relationship to a given diagnosis.

Successful probabilistic categorization requires learning the conditional probabilities (i.e., the likelihoods) of each feature occurring in each category, and the category base rates (i.e., their prior probabilities). According to decision theory [2], optimal categorization is achieved by taking, for each feature and combination, the product of likelihoods and priors to determine the *a posteriori* (posterior) probability of each category. When the goal is to be correct as often as possible, the optimal strategy is to always choose, for a given stimulus, the category that has the maximum *a posteriori* (MAP) probability [1, 3].

The present study asks, "What are the conditions that make a human decision maker more likely to adopt an optimal categorization strategy?" Categorization tasks can be of critical importance, as in medical diagnosis. The goal of this study is to provide a theoretical basis for understanding the conditions that not only serve to improve categorization accuracy but which serve to maximize it in such tasks.

The role of uncertainty regarding the MAP category. Having learned the posterior distribution over categories, the MAP category can be found by taking the mathematical expectation of the distribution. In the case of discrete categories, this amounts to calculating the posterior for each of the possible categories and picking the maximum. However, such full Bayesian inference, in which every possible category and its posterior are considered in turn, is intractable for human decision makers in all but the simplest of cases [4], and a body of research has shown that human decision making is inconsistent with such a strategy. People behave as if they base their decision on one or a small number of samples from the posterior rather than the expectation over the entire distribution [5].

The observation that people tend to base their decisions on a small number of samples drawn from the posterior distribution leads to a simple hypothesis regarding optimal probabilistic categorization: for a given stimulus, increasing the frequency with which it has been observed and/or the probability that it is in its MAP category will decrease the variance of the posterior (measured by, for example, its entropy [6]) around the MAP category. Subsequently, a sample drawn from the distribution will be more likely to indicate the MAP category. Our first hypothesis thus assumes that increasing the frequency and probability of a stimulus will decrease the observer's uncertainty as to its MAP category and they will

Significance Statement

In probabilistic categorization, each stimulus to be categorized can occur in each category with some probability. Here we examine the conditions that make a person likely to adopt a categorization strategy that maximizes their accuracy in such tasks. We show that adoption of the optimal strategy depends on the person's uncertainty with regard to the optimal decision rule and this uncertainty can be linked to 1) the relevant probabilities relating stimuli and categories, 2) the frequency with which the stimuli have been encountered in the process of learning the category relationships and, 3) the featural similarity of the stimuli. This work has relevance to understanding and improving performance and data presentation in real-world tasks such as medical diagnosis.

[1] The authors contributed equally to this work.

[2] To whom correspondence should be addressed. E-mail: jfackermann1@gmail.com

be more likely to consistently choose it.

The role of stimulus similarity. The hypothesis stated above assumes that the category membership of each stimulus is considered independently of that of all other stimuli. However, categorizing a stimulus has been shown to involve an assessment of its similarity to known members of the possible categories [7]. The decision as to category membership depends on stimulus similarity as well as probability and frequency. Changes in the probability of category membership of a stimulus can alter that of similar stimuli independent of their probabilities [8].

Stimulus similarity is subjective and based on the demands of the task [7, 9–11]. Dogs and cats, for example, are dissimilar if the task is to discriminate dogs from cats, but essentially identical if the task is to discriminate domesticated pets from automobiles. Similarity is generally modeled as a function of distance computed in the observer's subjective similarity space [12].

Our second hypothesis assumes that the structure of posterior distribution over categories depends on stimulus similarity in addition to their independent likelihoods and frequencies. It assumes that when presented with a set of stimuli to categorize, people will attempt to form a theory about the stimulus dimensions that are relevant to making a decision as to their category membership. Having arrived at a theory as to the relevant stimulus dimensions, people will simplify the decision by projecting the stimuli onto a 1-dimensional 'similarity space' in which similarity is a function of proximity in the space calculated using the appropriate distance metric. This accounts for the interdependence of similarity, probability, and frequency, as well as the task-dependent nature of stimulus similarity. The posterior distribution over categories is formed over the stimuli in similarity space [13]. We give a formal derivation and an example of this hypothesis below. But in sum, according to the second hypothesis, the posterior distribution over categories will depend on probability and frequency as well as stimulus similarity, and the degree to which people will adopt an optimal decision strategy will depend on category uncertainty (i.e., the entropy of the rule posterior) as in the first hypothesis.

The role of subjectively weighted probabilities. When people make decisions regarding events with probabilistic outcomes, they tend to behave as if they are using subjectively distorted versions of the relevant probabilities [14]. These subjective distortions typically take the form of a systematic overestimation of low probabilities and an under estimation of high probabilities [15]. Both of the models above assume that people will exhibit this typical pattern of probability weighting both with regards to the prior probabilities of the occurrence of each category and the observed proportions with which stimuli occur in each category. The way in which probability weighting is implemented within the models is discussed below.

The present study. In this study, we present two hypotheses regarding the factors that are conducive to adoption of an optimal categorization strategy. Both assume a primary role of the observer's ability to learn, for a given set of stimuli, a minimum-entropy representation of the posterior probability distribution over categories. The first hypothesis predicts a dependence on the magnitude of the MAP probabilities of the stimuli and the frequencies with which they have been encountered in the process of learning the association between stimuli and categories. The second predicts an additional dependence of a subjective measure of stimulus similarity. Both models assume that people's categorization decisions will exhibit typical patterns of subjective probability weighting.

We carried out an experiment with human subjects to examine the degree to which each of our two hypotheses predicts adoption of an optimal categorization strategy. Subjects were presented with stimuli consisting of three features, each of which could be present or absent, and placed each stimulus into one of two categories. The true category membership of the stimuli depended on independent probabilities associated with each of the three features thus making this a probabilistic task.

Prior to making category decisions, subjects completed a training phase in which examples of the stimuli were presented with their true category labels. We altered the MAP probability of the stimuli and the frequency with which each stimulus was presented in training in a way that allowed us to test for independent effects of probability and frequency on categorization in the testing phase.

We compared the human observers' performance to the predictions of two models representing our hypotheses and found that optimality is well predicted by a model representing the second hypothesis. Optimality is predicted by the subjects' uncertainty as to the relevant stimulus dimensions of the task and how the stimuli are best mapped onto subjective similarity space. Furthermore, we show that the source of this uncertainty can be found in the probabilities, frequencies, and similarity of the stimuli as calculated in their objective stimulus space.

Experimental Design. The experiment consisted of a training block in which subjects viewed stimuli with their true category labels followed by a testing block in which they categorized unlabeled figures. Stimuli consisted of the 3D figures shown in Fig. 1. Each stimulus consisted of 3 features, a top, middle, and bottom set of 'arms' that could be present or absent. The frequency and probability structure of the stimulus features were manipulated in the training block across 12 conditions. Each subject ran 1 of the training block conditions. All subjects ran the same testing block condition. We first introduce the frequency and probability structure of the testing block and then explain how it was modified to produce the 12 training block conditions.

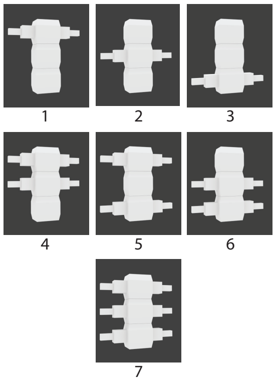

Fig. 1. The 7 stimuli used in the experiment.

Testing Block. In the testing block, each category had an equal probability of being the true category on a given trial. Let $p(C_i)$ equal the prior probability of category C_i (for $i = [1\ 2]$). Thus $p(C_1) = p(C_2) = .5$.

Each stimulus feature, i.e., the top, middle, and bottom set of arms, had an independent probability of occurring in each category. Let $p(F_j|C_i)$ for $j = [1\ 2\ 3]$ (representing the top, middle, and bottom set of arms, respectively) be the

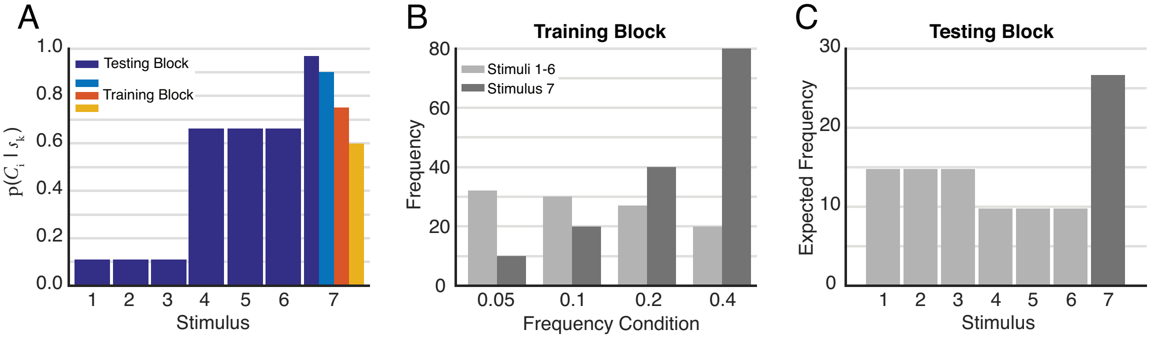

Fig. 2. Probaility and Frequency parameters of the experiment. **A** The posterior probability of Category 1 given each of the 7 stimuli in the Training and Testing Blocks. The probability of Stimulus 7 is varied across 3 conditions in the Training Block (see text). The probabilities for the other stimuli are identical in both blocks. **B** The frequency (number of trials) with which each stimulus is encountered in each of the 4 Frequency conditions (in proportion of trials for Stimulus 7) of the Training Block. **C** The expected frequency (number of trials) for each stimulus (N_k) in the Testing Block.

probability that feature F_j will occur in category C_i. In the testing block, $p(F_1|C_1) = p(F_2|C_1) = p(F_3|C_1) = .8$, and $p(F_j|C_2) = 1 - p(F_j|C_1)$.

On each trial, t, the true category, C^t, was calculated by a random draw from a Bernoulli distribution: $C^t \sim B\,(1, p(C_1))$. Having determined trial t's true category, the stimulus for that trial, $\mathbf{s}^t$, was created, where $\mathbf{s}^t = [s_1 \;\; s_2 \;\; s_3]$ is a vector indicating the presence or absence of each feature, F_j, such that $s_j = 1$ when F_j is present and 0 otherwise. The presence or absence of each feature was determined by 3 independent Bernoulli trials: $s_j \sim B\left(1, p(F_j|C^t)\right)$.

The probability that the k^{th} stimulus $\mathbf{s}_k$ will occur in category C_i is thus given by:

$$p(\mathbf{s}_k|C_i) = \frac{\prod_j p(s_{k,j}|C_i)}{\sum_k \prod_j p(s_{k,j}|C_i)} \qquad [1]$$

$$p(s_{k,j}|C_i) = \begin{cases} p(F_j|C_i) & s_{k,j} = 1 \\ 1 - p(F_j|C_i) & s_{k,j} = 0 \end{cases}$$

and the posterior probability that each category is true given each stimulus is:

$$p(C_i|\mathbf{s}_k) = \frac{p(s_k|C_i)p(C_i)}{\sum_i p(s_k|C_i)p(C_i)}. \qquad [2]$$

Using this probability structure results in a straightforward optimal decision rule for categorizing the stimuli: the MAP category of stimuli with more than 1 feature (stimuli 4, 5, 6, and 7) will be Category 1. The MAP category of stimuli with only one feature (stimuli 1, 2, and 3) will be Category 2.

The expected frequency with which the observer will encounter each stimulus in the Testing Block is determined by the probability structure: $N_k = N_T \sum_i p(s_k|C_i)p(C_i)$, where N_T is the total number of trials in the Testing Block. The posterior probabilities and expected frequencies for each stimulus in the Testing Block are shown in Fig. 2.

Training Block. In the Training Block, we manipulated stimulus probability and frequency to determine their effects on subsequent categorization performance. Specifically, we altered the probability of Stimulus 7, the stimulus consisting of all three features ($\mathbf{s_7} = [1\;1\;1]$) occurring in Category 1 across three conditions, $p(\mathbf{s_7}|C_1) = .60$, .75, and .90. We altered the frequency with which Stimulus 7 is observed in the training block, $N_{\mathbf{s_7}}$, over 4 conditions: $N_{\mathbf{s_7}} = 10$, 20, 40, and 80 trials (out of 200 trials total). The probabilities of stimuli $1 - 6$ occurring in each category were identical to those in the testing block. The frequencies of stimuli $1 - 6$ were equal in the training block, such that when, for example, Stimulus 7 occurred on 80 trials, stimuli $1 - 6$ occurred on 20 trials each of the remaining 120 trials. Note that altering the probabilities and frequencies in this way changed the category base rates in the Training Block. We examine the potential effects of this on observers' Testing Block responses in the Results section. The posterior probabilities and expected frequencies for each stimulus in the Training Block are shown in Fig. 2.

Predicted Effects. Under the first hypothesis, we would expect to observe an increase in the likelihood of adopting an optimal strategy for Stimulus 7 as its MAP probability and frequency increases in the Training Block. Likewise, as the proportion of trials on which Stimuli 1-6 are encountered in the Training Block decreases (see Fig. 2), Hypothesis 1 predicts a decrease in the likelihood of adopting an optimal strategy. If optimality depends only on the MAP probabilities and frequencies as in our first hypothesis, altering the probability of Stimulus 7 should have no effect on whether an optimal strategy is adopted for stimuli 1-6. If optimality depends on similarity we may find that increasing the category uncertainty of Stimulus 7 leads to a decrease in the likelihood that Stimuli 1-6 are categorized optimally.

Models

Probability-Frequency (PF) Model. The PF Model corresponds to our first hypothesis. It assumes that, in the Training Block, the observer counts the number of trials on which stimulus, $\mathbf{s}_k$, is seen in conjunction with the true category label, C_i:

$$N_{s_k|C_i} = \sum_{t=1}^{N_T} \left[\mathbf{s}_k^t \wedge C_i\right], \qquad [3]$$

where N_T is the total number of training trials, and $[x] = 1$ when x is true and 0 otherwise.

The observer derives an estimate of the probability that each stimulus, $\mathbf{s}_k$, will occur in each category:

$$\hat{p}(\mathbf{s}_k|C_i) = \frac{N_{s_k|C_i}}{\sum_{t=1}^{N_T} [\mathbf{s}_k^t]}. \qquad [4]$$

The observer's belief that stimulus $\mathbf{s}_k$ will occur in Category i is represented by the posterior probability given by:

$$p(C_i|\mathbf{s}_k) = \frac{w(\hat{p}(\mathbf{s}_k|C_i))w(p(C_i))}{\sum_i w(\hat{p}(\mathbf{s}_k|C_i))w(p(C_i))} \qquad [5]$$

$$w(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{\frac{1}{\gamma}}}, \quad [6]$$

where the prior probability $p(C_i)$ is made known explicitly to the observer in the instructions. $w(p)$ represents the subjective weighting that results in the overestimation of low and the under estimation of high probabilities [14]. Examples of $w(p)$ for various values of the parameter γ are shown in Fig. 3.

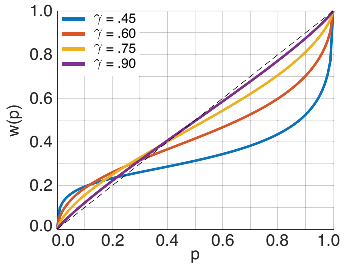

Fig. 3. The probability weighting function $w(p)$ for 4 values of the parameter γ.

The observer then forms a probability distribution over decision rules [16]. Each rule, R, is represented in the model by a logical formula in conjunctive normal form. For example, the optimal rule for the task is represented by: $(C_1) \iff (\mathbf{s}_4 \vee \mathbf{s}_5 \vee \mathbf{s}_6 \vee \mathbf{s}_7) \wedge \neg(\mathbf{s}_1 \vee \mathbf{s}_2 \vee \mathbf{s}_3)$,which reads, "Category 1 is true if and only if the current stimulus is 4, 5, 6, or 7 and not 1, 2 or 3." Each rule contains a positive followed by a negative disjunctive clause which we represent as $\mathcal{P}_{n,i}$ and $\mathcal{N}_{n,j}$, respectively, where n refers to the n^{th} rule and i and j refer to the i^{th} and j^{th} stimuli of the respective clauses. The n^{th} rule can thus be represented as:

$$R_n : (C_1) \iff \mathcal{P}_{n,i} \wedge \mathcal{N}_{n,j} \quad \forall i, j. \quad [7]$$

The rule posterior is given by the probability mass of a multinomial distribution over rules. The probability of the n^{th} rule given the training block stimuli and respective category labels is given by:

$$p(R_n | \mathbf{s}^1, ..., \mathbf{s}^{N_T}; C^1, ..., C^{N_T}) = \quad [8]$$

$$\frac{1}{c} \prod_i p(C_1 | \mathcal{P}_{n,i})^{N_{\mathbf{s}=\mathcal{P}_{n,i}|C_1}} \prod_j p(C_2 | \mathcal{N}_{n,j})^{N_{\mathbf{s}=\mathcal{N}_{n,j}|C_2}},$$

where c is the appropriate normalizing constant, and the N's represent stimulus counts as in Eq. 3.

There are 127 rules representing all possible parsings of the 7 stimuli into the rule clauses. On each trial of the Testing Block, the model draws a single sample from the rule posterior and chooses Category 1 or 2 with regard to that trial's stimulus according to the sampled rule.

Probability-Frequency-Similarity (PFS) Model. The PFS Model represents the second hypothesis. An illustration of it is shown in Fig. 4. The model assumes that the observer derives the posterior probabilities, $p(C_i | \mathbf{s}_k)$, as in Eq. 2. Then the observer will attempt to arrive at a theory regarding the stimulus dimensions relevant to categorization. For example, the observer may decide that the category posterior distributions indicate that (in keeping with the optimal decision rule) it is the number of features present in a stimulus that determines its category membership. Deciding to represent stimuli in terms of their number of present features effectively projects them onto a 1-dimensional space. Formally, if each stimulus is represented by the 3×1 vector $\mathbf{s}$ in stimulus space, where each element equals 1 if the feature is present and 0 otherwise, the observer would achieve the 1D projection by taking the inner product of each stimulus with the vector $[1\ 1\ 1]^T$ in stimulus space. The result is a scalar representation of each stimulus in 1D *similarity* space. Stimulus similarity is represented by the distance between the stimuli. For the case of the discrete-valued stimuli in this experiment, the appropriate distance metric is given by the absolute value of the scalar difference between stimuli in the 1D similarity space (i.e., the $L1$ difference). In the continuous case, the equivalent optimal projection vector is given by the first principle component of the category posterior distributions in stimulus space, and the appropriate distance metric is the L2 (Euclidean) distance.

Let $\mathcal{S}_d$ equal the set of all stimuli at the d^{th} value in the domain of the observer's subjective similarity space. Let $\mathbf{s}_{d,i}$ be the i^{th} stimulus in set $\mathcal{S}_d$. The observer derives an estimate of the probability that each set will occur in each category. The probability that set $\mathcal{S}_d$ belongs to Category 1 is:

$$p(\mathcal{S}_d | C_1) = \frac{\prod_i p(C_1 | \mathbf{s}_{d,i})^{N_{\mathbf{s}_{d,i}|C_1}}}{\sum_j \prod_i p(C_j | \mathbf{s}_{d,i})^{N_{\mathbf{s}_{d,i}|C_j}}}, \quad [9]$$

where the N's represent stimulus counts calculated as in Eq. 3.

The PFS Model assumes that uncertainty regarding the task relevant stimulus dimensions (and, hence, the best projection vector) may result in the mapping of the stimuli into fuzzy sets within similarity space such that each stimulus could be a member of multiple sets. Each set, $\mathcal{S}_d$, has an associated function, $m(\mathbf{s}_{d,i}) \in [0, 1]$, which gives the grade of membership of each stimulus in the set [17]. Equivalently, we define the function $m(\mathbf{s}_k, d)$ which gives the membership of each of the k stimuli in each of the d sets:

$$m(\mathbf{s}_k, d) = \frac{[\mathbf{s}_k \in \mathcal{S}_d]}{\sum_{j=1}^{N_D} [\mathbf{s}_k \in \mathcal{S}_j]}. \quad [10]$$

Thus $m(\mathbf{s}_k, d)$ equals 0 if $\mathbf{s}_k$ is not in set $\mathcal{S}_d$, and otherwise equals 1 divided by the number of sets of which stimulus $\mathbf{s}_k$ is a member.

The observer then forms a probability distribution over the decision rules, R_n. For each stimulus in the rule, the probability that it is in its respective category is the sum of the probabilities that each set is in the category weighted by the membership of the stimulus in each set. The rule posterior is then the product of these probabilities:

$$p(R_n | \mathcal{S}, m) = \quad [11]$$

$$\frac{1}{c} \prod_i \sum_d m(\mathcal{P}_{n,i}, d) p(\mathcal{S}_d | C_1) \prod_j \sum_d m(\mathcal{N}_{n,j}, d) p(\mathcal{S}_d | C_2),$$

where c is the appropriate normalizing constant.

As in the PF Model, a single sample is drawn from the rule posterior on each trial of the Testing Block, and Category 1 or 2 is chosen with regard to that trial's stimulus according to the sampled rule. Implementation of both models is discussed in the Results section.

Results

20 subjects participated in each of the 12 Training Block conditions. Data consist of each subject's category choices in the Testing Block.

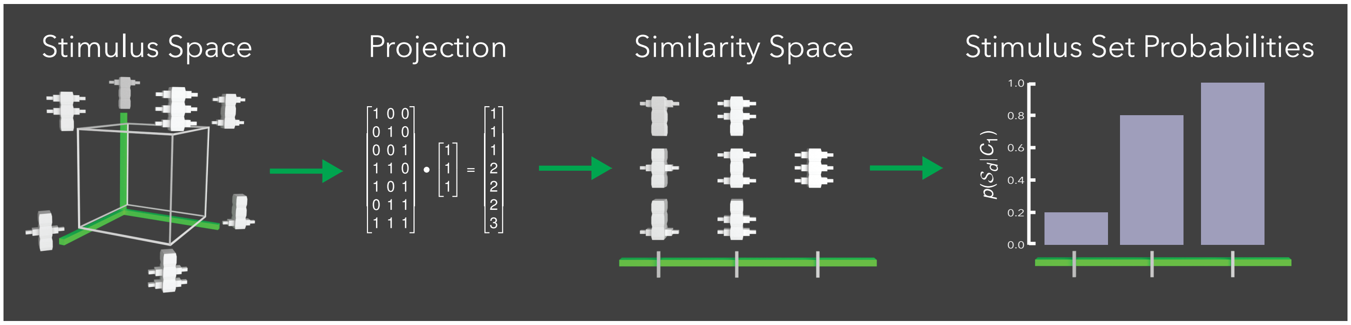

Fig. 4. The PFS Model. The 7 stimuli used in the experiment can be represented as vectors that lie at the corners of a cube in a 3D stimulus space. In order to categorize the stimuli, observers will form a theory regarding the relevant stimulus dimensions. Deciding on the relevant dimensions effectively projects the stimuli onto a 1-dimensional Similarity Space. For example, if the observer believes that it is the number of features possessed by the stimulus that determines its category membership, this results in the projection of each of the 3D stimulus vectors (each element of which equals 1 if the feature is present and 0 otherwise) onto the vector representing the presence of all 3 features. The result is a scalar-valued representation of the stimuli in terms of their location in Similarity Space. Next the observer derives distributions over Similarity Space representing the probabilities that the set of stimuli at each point in the domain of the space will occur in each category and then uses these probabilities to form the posterior distribution over possible decision rules (see text).

Response Bias. 33 out of 240 subjects had a significant bias for choosing one category over the other. Significance was determined by χ^2 test ($\chi^2(1)$ ranged from 3.92 to 12.5; p values ranged from .0004 to .048). These subjects were excluded from all subsequent analyses. A full analysis of response bias is given in the Supporting Information (SI).

Categorization Accuracy. Next we investigate whether manipulating stimulus probability and frequency in the Training Block conditions had a significant effect on subjects' categorization accuracy in the Testing Block. We calculated each subject's accuracy as the proportion of correct categorizations for each stimulus and Training Block condition. Median accuracy across subjects for each stimulus and Training Block condition is shown in Fig. 5.

We tested for significant effects on accuracy using a $3 \times 4 \times 7$ mixed-design ANOVA (i.e., 3 probability conditions by 4 frequency conditions by 7 stimuli in which *stimulus* is a within-subjects variable). There are significant main effects on accuracy of probability condition ($F(2, 195) = 3.54, p = .033$) and stimulus ($F(6, 1170) = 49.1, p < .001$) and significant interactions between probability condition and stimulus ($F(12, 1170) = 2.82, p < .001$) and frequency condition and stimulus ($F(18, 1170) = 1.7, p = .033$).

The presence of significant interactions makes interpretation of the main effects ambiguous. The main effect of probability condition on accuracy may merely reflect the overall increase in accuracy for categorizing Stimulus 7 as its MAP probability is increased across the 3 Training Block conditions (Fig. 5 bottom panel). It may also reflect an overall increase in accuracy across stimuli as can be seen in Fig. 6 A. The main effect of stimulus is likely reflected by the lower overall accuracy for categorizing stimuli 4, 5, and 6. This lower accuracy is to be expected since the MAP probabilities of stimuli 4, 5, and 6 are closer to .5 than those for the other stimuli making the maximum achievable accuracy (i.e., when the MAP category is chosen on every trial) for these stimuli lower (Fig. 6 C). Although the effect of frequency condition was not significant (Fig. 6 B), the interaction between frequency condition and stimulus may reflect the decrease in accuracy for categorizing Stimulus 7 as its frequency is increased across the Training Blocks (Fig. 5 bottom panel). This trend is counter to the prediction above that increasing stimulus frequency should decrease its category uncertainty and result in an *increase* in accuracy.

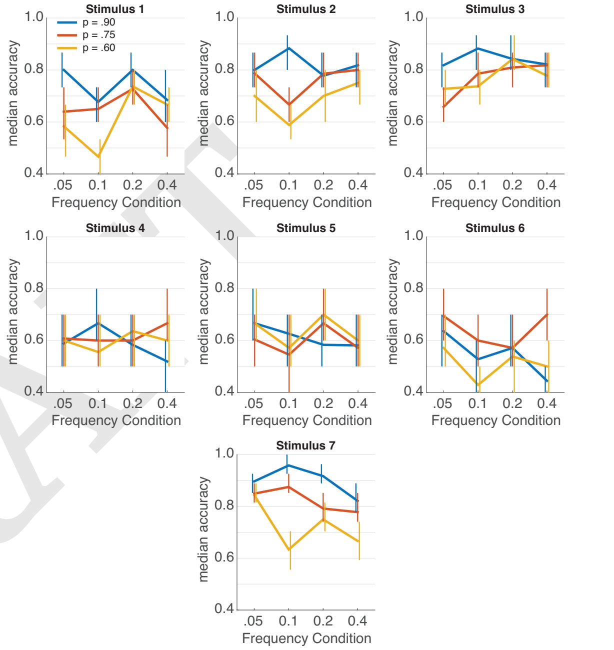

Fig. 5. Median categorization accuracy as a function of Training Block frequency condition across subjects for each stimulus. Each line represents one of the three probability conditions in the Training Block. Error bars show the central 50^{th} percentile of a normalized binomial distribution with N equal to the expected frequency of each stimulus (as in Fig. 2C).

Categorization Optimality. Next we look for significant effects of the probability and frequency conditions on subjects' tendency to adopt an optimal categorization strategy. Let $p(C^{MAP}|\mathbf{s}_k)$ equal the proportion of trials on which a subject chose the MAP category for stimulus, $\mathbf{s}_k$. The subject adopted an optimal categorization strategy if they chose C^{MAP} on every trial that stimulus $\mathbf{s}_k$ was presented and thus optimality is indicated when $p(C^{MAP}|\mathbf{s}_k) = 1$. The proportion of subjects that adopted an optimal categorization strategy for each stimulus, probability condition, and frequency condition is shown in Fig. 7.

To determine whether the Training Block conditions had a significant effect on subjects' tendency to choose the MAP categories in the Testing Block, we calculated a $3 \times 4 \times 7$ mixed-design ANOVA with $p(C^{MAP}|\mathbf{s}_k)$ as the dependent variable. We find significant main effects of probability condi-

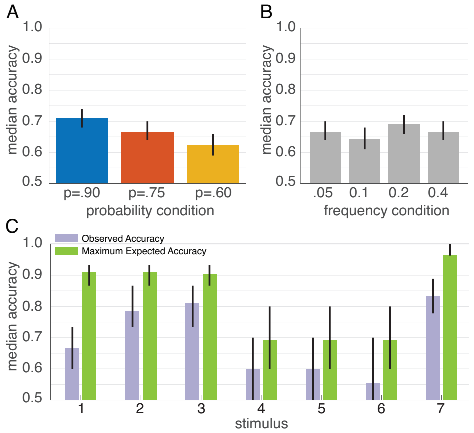

Fig. 6. **A** Median categorization accuracy across subjects for each Training Block probability condition, **B** for each frequency condition and, **C** observed accuracy and the maximum expected accuracy (see text) for each stimulus. Error bars show the 50^{th} percentile of a binomial distribution over proportion correct.

tion ($F(2, 195) = 4.27, p = .015$) and stimulus ($F(6, 1170) = 20.19, p < .001$). The interactions between probability condition and stimulus, and frequency condition and stimulus are not significant ($F(18, 1170) = 1.145, p = .32$, and $F(12, 1170) = 1.48, p = .088$, respectively). We tested for significant differences between the 3 probability condition levels and the 7 stimuli using 2-tailed 2-sample t-tests. Corrected α values for the multiple comparison were calculated by minimizing the false detection rate using the Benjamini-Hochberg method [18]. The results are shown in Fig. 8.

The main effect of probability condition on optimality reflects the greater proportion of subjects that adopted an optimal categorization strategy in probability condition 1 when the probability of Stimulus 7 occurring in Family 1 was .9 (Fig. 8 A). This effect is of interest since it indicates that changing the MAP probability of Stimulus 7 affected the categorization choices for the other stimuli whose MAP probabilities were unchanged. The greatest proportion of subjects adopted the optimal categorization strategy for Stimulus 3 and this proportion was significantly higher than that for Stimuli 1 and 5 (Fig. 8 C).

Finally, we ask whether a stimulus' MAP probability in training and the frequency with which it is encountered in training has an effect on choosing the MAP category in the Testing Block *independent of the probability and frequency conditions of the Training Block*. We calculated a 2-way repeated-measures ANOVA for which the dependent variable is $p(C^{MAP}|\mathbf{s}_k)$ and the independent variables are, for each stimulus, the probability that it will occur in its MAP category in training and the frequency with which it was presented in training. The analysis shows no main effects for either probability or frequency. Fig. 9 shows the proportion of trials on which the MAP category was chosen for each stimulus as a function of its MAP probability (left panel) and frequency (right panel).

Summary. The effects of the probability and frequency conditions on categorization accuracy depended on an interaction with the stimuli. As predicted by both hypotheses, subjects had higher accuracy for categorizing Stimulus 7 when its MAP probability in the Training Block was higher. But the effect of probability condition may also reflect an overall increase in accuracy for all stimuli when Stimulus 7's MAP probability

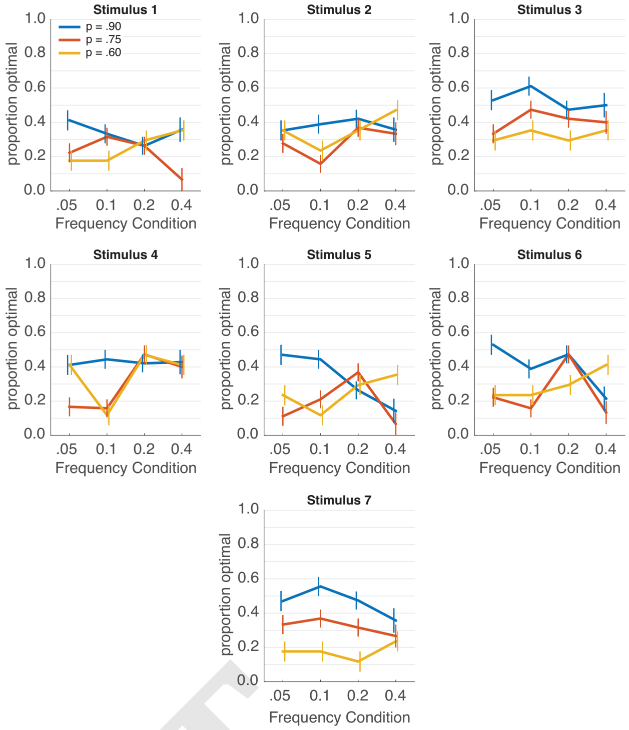

Fig. 7. Proportion of subjects who adopted the optimal categorization strategy for each stimulus and each Training Block condition. Error bars show the 50^{th} percentile of a binomial distribution with N equal to the number of subjects in each condition.

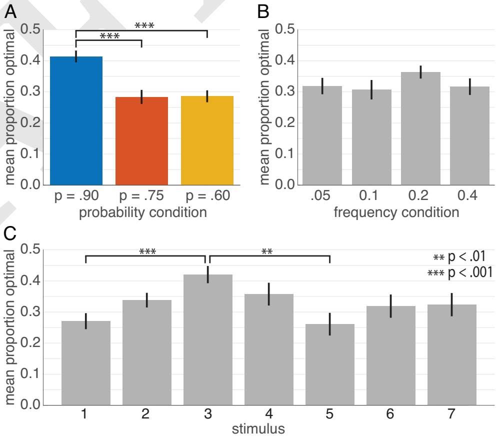

Fig. 8. Weighted average proportion optimal **A** for each Training Block probability condition, **B** each frequency condition and, **C** each stimulus. The weights reflect the different numbers of subjects in each condition. The error bars show 1 SE above and below the mean.

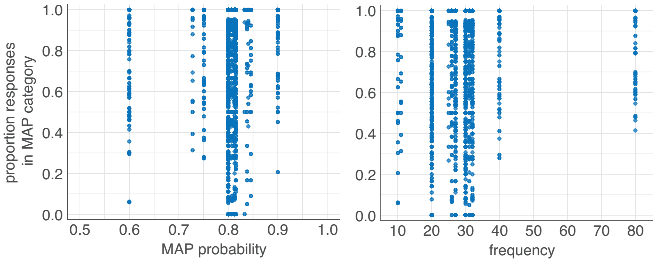

Fig. 9. Proportion of responses in the MAP category as a function of stimulus MAP probability (Left panel) and frequency (Right panel) in the Training Block. Each dot represents 1 stimulus viewed by a signal subject. The same data appear in each panel. The MAP probability and frequency of the independent stimuli had no effect on subjects' tendency to assign them to their MAP category (see text).

was altered as predicted by the second hypothesis. Categorization accuracy for Stimulus 7 shows a decreasing trend as its frequency is increased across the Training Block conditions. This is counter to Hypothesis 1.

The effect of probability condition on optimality is less ambiguous. In keeping with Hypothesis 2, a greater proportion of subjects adopted the optimal categorization strategy for all stimuli when Stimulus 7's MAP probability in the Training Block was .9. And the individual MAP probabilities and frequencies of each stimulus in the Training Block had no effect on the subjects' propensity to choose its MAP category in the Testing Block.

Model Predictions.

PF Model. PF Model predictions of categorization accuracy and optimality were generated for each subject by first deriving the posterior distribution over decision rules as given by Eq. 8 where the stimulus vectors, $\mathbf{s}$, and category labels, C, were those seen by the subject in the Training Block. For each trial in the Testing Block, the model drew a single sample from the rule posterior and responded to the respective stimulus according to the sampled rule. For each subject, 1000 iterations of the simulated experiment were run. For each iteration, we calculated the number of trials on which each category was selected for each stimulus. From these values we derived the binomial likelihood of the simulated runs given the proportions with which the subjects' selected each category for each stimulus. The optimal value of the weighting function parameter γ (Eq. 6) was fit to the data by maximizing this likelihood. From the 1000 iterations derived using the optimal value of γ, for each stimulus in each condition, we then take the proportion of simulated runs on which the model adopted the optimal categorization strategy. That is, the proportion of simulated runs on which the model chose the MAP category for all 100 trials of the Testing Block. The median proportion of optimal responses made by the PF Model, across subjects and for each stimulus in each condition, is shown in Fig. 10.

The PF Model's predictions diverge from the data in two principle ways: first, the PF Model predicts an increase in optimality for Stimulus 7 as its frequency increases. Second, it predicts an overall decrease in optimality for Stimuli $1 - 6$ as their relative frequency decreases. These trends are not evident in the subjects' data (Fig. 10).

In each probability condition, as the frequency of Stimulus 7 increases, the model predicts an increasing proportion of optimal responses. The subjects' data show the opposite pattern with optimal responses decreasing with the increase in frequency. Similarly, as the frequency of Stimulus 7 increases, the frequencies of stimuli 1 - 6 decrease and the model predicts a decrease in the proportion of optimal responses for stimuli 1 - 6. The subjects' data do not show this consistent decrease and in fact there is an increases in proportion optimal for stimuli 1 - 6 across the frequency conditions in many cases. Since the MAP probabilities of stimuli 1 - 6 do not differ across the three probability conditions, Model 1's predicted proportion optimal does not differ across the probability conditions for stimuli 1 - 6. Subjects' proportion optimal, however, shows substantial differences across the probability conditions for stimuli 1 - 6 consistent with the main effect of probability condition found in the above analysis.

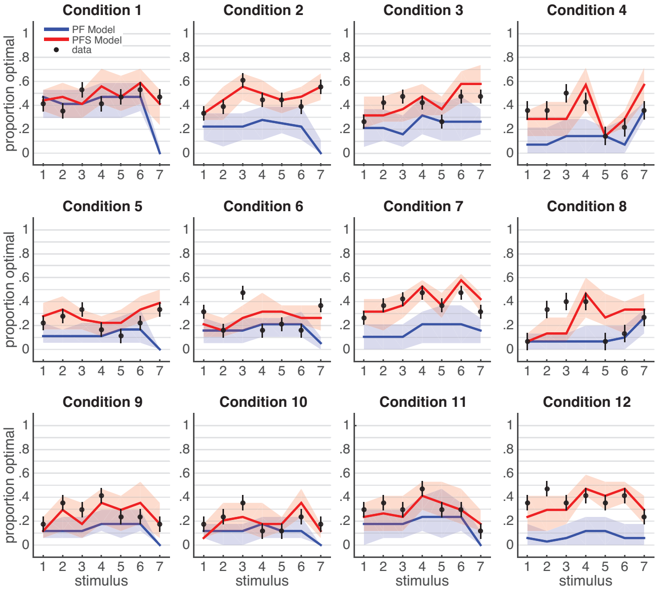

Fig. 10. The proportion of subjects who adopted the optimal categorization strategy for each stimulus following each of the 12 Training Block conditions. Dots show the subjects' data as in Fig. 7. Lines show the median predicted proportion optimal of each model. Shaded regions show the central 95^{th} percentile of the predictions.

PFS Model. The PFS Model assumes that subjects attempt to map the stimuli onto a 1-dimensional similarity space (Fig. 4). In order to derive the PFS Model predictions, it is necessary to know each subject's stimulus mapping as represented by $\mathcal{S}$ (Eq. 9). This information can only be inferred from their data and thus, in order to implement the model, the stimulus mappings, $\mathcal{S}$, (and the optimal value of the probability weighting parameter γ) were fit to the subject's responses in the Testing Block using a Genetic Algorithm (GA). While it is possible, although impractical, in the case of the 3-feature stimuli used in this experiment to check the goodness of fit of every possible stimulus mapping (there are 7^7 possible mappings), if the number of features is 4 or greater this becomes intractable, and thus the GA was selected as a fitting method amenable to use with stimuli of any dimension.

In implementing the GA, the estimates of $\mathcal{S}$ were derived by directly fitting the membership function, $m(\mathbf{s}_k, d)$ (Eq. 10). The best fitting estimate of $\mathcal{S}$ was selected by simulating (on each iteration of the GA) the performance of an observer that uses $\mathcal{S}$ to derive the posterior distribution over decision rules using Eq. 11 and then making category decisions for all stimuli viewed by the subject as described above. The GA's estimates of $\mathcal{S}$ and γ were optimized by maximizing the likelihood of the simulated category choices with respect to the subject's actual choices. A full description of the implementation of the PFS Model, the predicted stimulus mappings, $\mathcal{S}$, and the predicted category choices of each model for each subject is given in the SI.

To derive the proportion of optimal responses for the PFS Model, we simulated the experiment as with the PF Model and calculated the proportion of simulated runs on which the model adopted an optimal categorization strategy. The median proportion, across subjects, of optimal responses by the PFS Model, for each stimulus in each condition, is shown in Fig. 10.

The predictions of the PFS Model provide a qualitatively better fit to the data although, in Conditions 4 and 8, it tends to underestimate optimality for stimuli 2 and 3 and overestimate that for Stimulus 7. We discuss these failures of the model below.

We quantify the difference between model predictions and observers' performance by asking whether, given a particular set of stimuli and training conditions, each model will provide a prediction that is not significantly different from the observers' performance. Thus we test for a significant difference across stimuli for each of the 12 conditions (3 probability by 4 frequency) of the experiment. We determined significance as follows: the model simulations result in sets of the number of subjects predicted to adopt the optimal decision strategy for each stimulus in each condition. We assume these sets of frequencies to be binomially distributed and fit distributions to each set. We then derive an empirical likelihood distribution for each set by calculating the likelihood of each simulated frequency in the set. For each of 10,000 runs of the model simulations, we sum the likelihood of the predicted number of subjects adopting optimal strategies across the stimuli within each condition. We then determined the upper 95^{th} percentile of this empirical likelihood distribution for each condition. We derived the likelihood of the observed number of subjects who adopted an optimal strategy across the stimuli in each condition and compared these values to the 95^{th} percentiles of the empirical likelihoods for each condition. The observed proportions differ significantly from Model 1 in all conditions and from Model 2 in Conditions 4, 8, and 12 . As mentioned above, the significant difference with respect to the PFS Model likely reflects the consistent underestimation of proportion optimal for stimuli 2 and 3 in these conditions as well as the over estimation for Stimulus 7 in condition 4 (see Fig. 7). The empirical likelihood distributions and observed likelihoods given each model are provided in the SI.

Next we determine the overall ability of each model to predict the proportion of optimal subjects for all stimuli in all conditions. We carried out the Total Least Squares Regression of the predicted proportion of optimal subjects for each model with respect to the observed proportion optimal. The results are shown in Fig. 11. The regression slopes of each model are not significantly different from 1 as determined by two-tailed t-test. The PF Model tends to underestimate proportion optimal with a slope of .71 and has an r^2 value of .11. Model 2's regression slope is 1.08 with an r^2 value of .49.

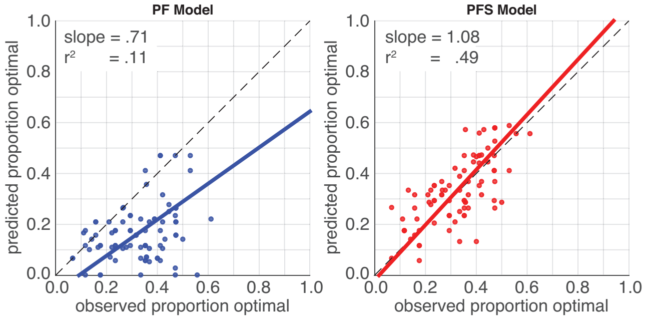

Fig. 11. Predicted as a function of observed proportion optimal for the PF (Left panel) and PFS (Right panel) models. Each dot represents 1 stimulus in 1 Training Block condition. Solid lines show the TLS regression lines (see text).

Probability weighting function parameters. In order to ensure that subjects' fit probability weighting function parameters, γ, did not vary dependent on experimental condition, we calculated a 1-way ANOVA comparing each model's fit values of γ across the 12 conditions. We found no significant effect of experimental condition on the value of γ for either model (PF Model: mean = .83, SD = .08; PFS Model: mean = .86, SD = .07). In addition, we expect each subject's two estimated values of γ (one for each model) to be significantly correlated. We calculated Pearson's ρ for the correlation of γ estimates for each model. The resulting $\rho = .23$ and the correlation is significant ($t(205) = 3.32$, $p = .0011$). A full analysis of the γ values is included in the SI.

Predicting optimality using *Relative Uncertainty*. The PFS Model suggests that a suboptimal categorization strategy will arise when the observer is uncertain as to the optimal decision rule, i.e., when the entropy of the posterior probability distribution over rules is high. Furthermore, rule-based uncertainty stems from uncertainty regarding how the stimuli are to be mapped onto a subjective similarity space. But the PFS Model does not offer an explanation as to the source of this stimulus-mapping uncertainty. Here we derive a measure that attempts to locate the source of uncertainty leading to suboptimal categorization within the relevant probabilities, frequencies, and 'objective' similarity of the stimuli to-be-categorized.

Let $P_{k,i} = p(C_i|\mathbf{s}_k)$ equal the posterior probability (Eq. 2) that category C_i is true given the stimulus, $\mathbf{s}_k$. And let the *evidence* of $P_{k,i}$ equal $P_{k,i}^{N_{k,i}}$, where $N_{k,i} = N_{s_k|C_i}$ are stimulus counts calculated as in Eq. 3. The standard deviation of the proportion $P_{k,i}$ is given by $\sigma_{k,i} = \sqrt{P_{k,i}(1-P_{k,i})/N_{k,i}}$. And we define the *standard error of the evidence* as $\sigma_{k,i}/\sqrt{N_{k,i}}$. Finally, we define the matrix $\mathbf{D}_{K,K}$ to represent the similarity between each of the K stimuli in the data set. Each element of $\mathbf{D}_{k,n}$ (for $k, n = \{1, ..., K\}$) equals $\frac{1}{1+\Delta_{k,n}}$, where Δ is the appropriate distance metric, in this case, the $L1$ distance given by:

$$\Delta_{k,n} = \sum_j |\mathbf{s}_{k,j} - \mathbf{s}_{n,j}|, \qquad [12]$$

where the sum is over the j dimensions of the stimuli.

We define the *weighted evidence*, E, for the k^{th} stimulus occurring in the i^{th} category as the sum of the evidence of each stimulus in the set, where each stimulus' evidence is weighted by the reciprocal of its standard error and its similarity with regards to stimulus k:

$$E_{k,i} = \sum_n \mathbf{D}_{k,n} \frac{\sqrt{N_{n,i}} P_{n,i}^{N_{n,i}}}{\sigma_{n,i}}. \qquad [13]$$

The Relative Uncertainty, U, of the dataset is defined as the reciprocal of the sum of the weighted evidence across stimuli and categories:

$$U = \frac{1}{\sum_i \sum_k E_{k,i}}. \qquad [14]$$

U thus represents a measure of the overall uncertainty associated with making a category decision given the data. Mean proportion optimal as a function of U for each of the 12 conditions of the experiment is plotted in Fig. 12. Proportion optimal appears to follow an exponential decay function of U. The best fitting least squares fit: $.46e^{-8.27U}$ is plotted in red.

It can be seen from Eqs. 13 and 14 that relative uncertainty, U, will decrease as the MAP probabilities, P, and/or the frequencies, N, increase. In addition, U will decrease as stimulus similarity is increased (as predicted by Yager [19]). An example of this effect is given in the SI.

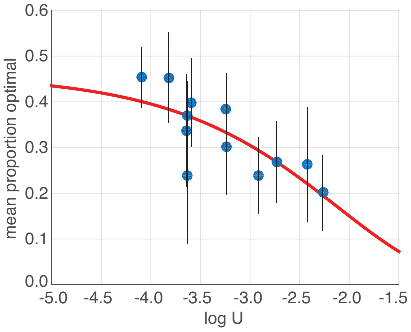

Fig. 12. Mean proportion optimal for each of the 12 Training Block conditions as a function of log Relative Uncertainty, U. Each dot represents 1 condition. Error bars show 1 SE above and below the mean. The solid line shows the best fitting exponential decay function (see text).

Discussion

In this study, we investigated the conditions that make a person more or less likely to adopt an optimal decision strategy in a probabilistic categorization task. The optimal strategy in this context is the one that maximizes accuracy and involves choosing the MAP category for a given stimulus every time it is encountered. Optimal probabilistic categorization is thus achieved by learning an accurate representation of the posterior probability distributions over stimuli. Our hypotheses make the specific assumption that an optimal strategy is more likely to be adopted when the observer has learned a minimum-variance representation of the category posteriors and hence of the posterior distribution over decision rules. Our first hypothesis, represented by the PF model, posits that this depends critically on the magnitude of the learned posteriors and the frequency with which each stimulus has been encountered in the course of learning. Our second hypothesis, represented by the PFS model, assumes an additional role of stimulus similarity and that adoption of the optimal strategy for a given stimulus depends on a subjective measure of its similarity to all stimuli in the relevant feature space.

The results of our experiment suggest a role of probability, frequency, and similarity as represented by the PFS model. The predictions of the PFS model are not significantly different from subjects' observed proportion optimal in 9 out 12 conditions. We discuss the failures of the model below. Furthermore we show that a measure of categorization uncertainty, U given in the Eq. 14, is a good predictor of changes in proportion optimal that occur in association with changes in the relevant probabilities and frequencies as well an objective measure of stimulus similarity. This measure thus relates observers' decision uncertainty to measurable aspects of the stimuli to-be-categorized and may serve as a useful tool in predicting changes in optimality between different data sets and between stimulus representations for a given data set.

Limitations of the present study and future directions.

Failures of the PFS model. The PFS model's predictions differed significantly from observed proportion optimal in Conditions 4, 8 and 12 (Fig. 10). In these conditions, the frequency of Stimulus 7 in the Training Block was high (seen on 80 trials) compared to the other stimuli (20 trails each). Stimulus 7's MAP category is Category 2. In short, the PFS model fails to account for subjects' tendency to choose Category 1 for Stimulus 7 in these conditions when it is most likely to occur in Category 2 and has a high frequency. It assumes that Stimulus 7 is associated with Stimuli 1, 2, and 3, i.e., stimuli which have a high probability of occurring in Category 1. This association raises the PFS model's estimate of Stimulus 7 responses in Category 1, but also raises the estimates of Stimuli 1, 2, and 3 responses in Category 2. The overall result is an overestimation of proportion optimal for Stimulus 7 and an underestimation for Stimuli 1, 2, and 3. Note, though, the proximity of the PF model's predictions to Stimulus 7's observed proportion optimal, particularly in Conditions 4 and 8 (Fig. 10). The PF model treats each stimulus independently and chooses categories without reference to a measure of similarity. Its predictions of proportion optimal for Stimulus 7, particularly for Conditions 4 and 8, appear accurate. We speculate that observers may have adopted an independent decision rule for Stimulus 7 in these conditions. Future research may explore this hypothesis that high-frequency, high-probability stimuli are categorized independently of other stimuli in the feature space.

Simplistic priors, likelihoods, and utilities. In this study, we have examined categorization in the simplest case in which, 1) the category prior probabilities, $p(C_i)$, are equal, 2) the probabilities of each feature occurring in each category are independent, and 3) performance is optimized by maximizing accuracy. Probabilistic categorization can be subject to base rate neglect [20, 21] such that people seem to make decisions without regard to the priors when they differ. Future research may examine whether base rate neglect is observed in the present study's paradigm and, if so, whether the PFS model's predictions remain accurate. Insofar as it relies on normative Bayesian statistics (Eq. 2), the PFS model in its current state may fail to account for such behavior.

Our models calculate the joint likelihoods of the features co-occurring in each stimulus assuming that they do not co-vary (Eq. 1). Increasing feature covariance has been shown to decrease category discriminability [22]. Future research may examine the inclusion of non-zero covariance in the PFS model for the dependent case in which some features are more likely to co-occur than others. This would represent a more realistic scenario as when, for example, some medical symptoms are more likely to indicate a positive diagnosis when seen in conjunction with certain other symptoms.

The experimental task used in this study had no concrete rewards or penalties associated with the subjects' responses. Subjects who adopted the optimal categorization strategy optimized the simplistic utility function in which a correct response represents a gain and there are no potential losses. The imposition of rewards and penalties has been shown to affect decision criteria in categorization tasks [23–25]. Attaching a penalty to an incorrect response can alter the proportions with which the observer assigns the stimuli to the respective categories. Future research may examine the PFS model's predictions in tasks with concrete rewards and penalties. The PFS model assumes that the observer's stimulus mapping, $\mathcal{S}$, is task-dependent and chosen to be most useful in correctly categorizing the stimuli. The imposition of a given observer's utility function for gains and losses [14] will likely result in a different mapping for the same stimuli, probabilities, and frequencies, that serves to increase gain. The measure of *weighted evidence*, E (Eq. 13), may require a term representing the given stimulus' associated reward and penalty.

The role of irrelevant stimulus properties. Finally, we have asserted that in examining optimal categorization behavior, it is necessary to consider the relevant probabilities, frequencies, and stimulus similarity. But these are not the only factors that can affect optimal categorization. The salience of a feature can have an effect, where salience is some factor that increases its relative detectability or discriminability [26], regardless of its likelihood of occurring in a given category. In addition, the presence of uninformative, irrelevant features, can affect categorization. Both salient and irrelevant features have the effect of causing other features to be 'underutilized' or ignored to a degree in making categorization decisions [27] and the presence of irrelevant features can affect learning of the relationship between categories and the relevant stimulus dimensions [28]. The effect of salience and irrelevance is generally modeled as a weight applied to the measure of distance between pairs of stimuli in similarity space [10, 26] which effectively stretches and compresses the space in response to attentional allocation to certain stimulus dimensions. In the context of the PFS model, this corresponds to a different (e.g., continuous) parameterization of the membership function, $m(\mathbf{s}_k, |d)$. Future research may examine the role of irrelevant stimulus properties on optimality and whether they are accounted for by distortions of similarity space in this way.

Materials and Methods

Participants. 240 subjects (median age = 31; range = 19 - 59 yrs) participated using Amazon's Mechanical Turk. All workers were based in the U.S. and had a 95% or greater approval rating. Of 151 subjects who reported their gender, 53% were female. All indicated their informed consent according to the guidelines of Northeastern University's IRB. The experiment was carried out in a single session. Median run time was 11 minutes. Subjects received $1.66 for their participation.

Stimuli. Each figure consisted of a top, middle, and bottom segment. Each segment could include a set of horizontally oriented 'arms'. Each figure had at least one set of arms. Stimuli were rendered using the Blender 3D rendering package (blender.org). Experiments were created using Javascript/HTML5 and the jsPsych Javascript library [29].

Experimental Sequence. Subjects first viewed a consent document and indicated their informed consent by mouse click. They then began the experiment consisting of 5 blocks run in sequence:
1) *Training Instructions.* In the Training Instruction block, subjects were introduced to the stimuli and told that they would be learning to assign each 'figure' to one of two 'families'. The full text of the instructions is available in the SI. Most importantly, subjects were instructed A) that each figure could occur in each family but would be more common in one family than the other, and B) that approximately half of the figures were from 'family 1' and half from 'family 2'. They were instructed that they would be shown a set of figures with their respective family labels and that they should try to learn which figures occur in which family.
2) *Training Block.* Subjects then completed 200 trials of the Training Block. On each trial, stimulus $\mathbf{s}^t$ and category label C^t were visible (Fig. 13A). The screen remained visible until the subject advanced to the next trial by mouse click.
3) *Testing Instructions* This block consisted of a single screen instructing the subject that they would next view a series of figures and decide which family each is from.
4) *Testing Block.* Subjects completed 100 trials on which stimulus $\mathbf{s}^t$ was visible (Fig. 13B) They indicated their decision as to which family the stimulus was from by clicking on label '1' or '2'. The label indicating their choice changed color briefly before automatically advancing to the next trial. The subject had unlimited time to complete each trial. No feedback was given.
5) *End Survey.* Subjects were shown a single page on which they were asked to indicate their age, gender, and any patterns they noticed or strategies they used to complete the task.

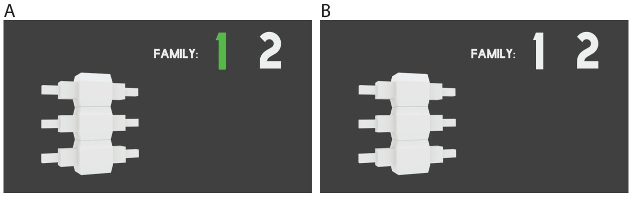

Fig. 13. Example stimulus screen viewed by the subject in **A** the Training Block and, **B** the Testing Block. See text for description.

1. Estes W (1972) Elements and patterns in diagnostic discrimination learning. *Transactions of the New York Academy of Sciences* 34(1 Series II):84–95.
2. Coombs CH, Dawes RM, Tversky A (1970) *Mathematical psychology: An elementary introduction.* (Prentice-Hall).
3. Kubovy M, Healy AF (1977) The decision rule in probabilistic categorization: What it is and how it is learned. *Journal of Experimental Psychology: General* 106(4):427.
4. Kwisthout J, Wareham T, van Rooij I (2011) Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science* 35(5):779–784.
5. Vul E, Goodman N, Griffiths TL, Tenenbaum JB (2014) One and done? optimal decisions from very few samples. *Cognitive science* 38(4):599–637.
6. Lindley DV (1956) On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics* pp. 986–1005.
7. Goldstone RL (1994) The role of similarity in categorization: Providing a groundwork. *Cognition* 52(2):125–157.
8. Estes WK (1986) Array models for category learning. *Cognitive psychology* 18(4):500–549.
9. Nosofsky RM (1984) Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, and cognition* 10(1):104.
10. Nosofsky RM (1986) Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General* 115(1):39.
11. Medin DL (1989) Concepts and conceptual structure. *American psychologist* 44(12):1469.
12. Medin DL, Schaffer MM (1978) Context theory of classification learning. *Psychological review* 85(3):207.
13. Ashby FG, Alfonso-Reese LA (1995) Categorization as probability density estimation. *Journal of mathematical psychology* 39(2):216–233.
14. Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society* pp. 263–291.
15. Fox CR, Poldrack RA (2009) Prospect theory and the brain in *Neuroeconomics: Decision making and the brain*, eds. Glimcher PW, Fehr E. (Elsevier London), pp. 145–174.
16. Goodman ND, Tenenbaum JB, Feldman J, Griffiths TL (2008) A rational analysis of rule-based concept learning. *Cognitive Science* 32(1):108–154.
17. Zadeh LA (1965) Fuzzy sets. *Information and control* 8(3):338–353.
18. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* pp. 1165–1188.
19. Yager RR (1992) Entropy measures under similarity relations. *International Journal Of General System* 20(4):341–358.
20. Gluck MA, Bower GH (1988) From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General* 117(3):227.
21. Estes WK, Campbell JA, Hatsopoulos N, Hurwitz JB (1989) Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15(4):556.
22. Alfonso-Reese LA, Ashby FG, Brainard DH (2002) What makes a categorization task difficult? *Attention, Perception, & Psychophysics* 64(4):570–583.
23. Maddox WT, Bohil CJ (1998) Base-rate and payoff effects in multidimensional perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24(6):1459.
24. Maddox WT, Dodd JL (2001) On the relation between base-rate and cost-benefit learning in simulated medical diagnosis. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27(6):1367.
25. Maddox WT (2002) Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the experimental analysis of behavior* 78(3):567–595.
26. Kruschke JK, Johansen MK (1999) A model of probabilistic category learning. *Journal of Experimental Psychology Learning Memory and Cognition* 25:1083–1119.
27. Edgell SE et al. (1996) Irrelevant information in probabilistic categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22(6):1463.
28. Little DR, Lewandowsky S (2009) Beyond nonutilization: Irrelevant cues can gate learning in probabilistic categorization. *Journal of Experimental Psychology: Human Perception and Performance* 35(2):530.
29. De Leeuw JR (2015) jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods* 47(1):1–12.