Feature inference with uncertain categorization: Re-assessing Anderson's rational model

Elizaveta Konovalova¹ · Gaël Le Mens^{1,2}

Published online: 18 September 2017
© Psychonomic Society, Inc. 2017

Abstract A key function of categories is to help predictions about unobserved features of objects. At the same time, humans are often in situations where the categories of the objects they perceive are uncertain. In an influential paper, Anderson (*Psychological Review*, 98(3), 409–429, 1991) proposed a rational model for feature inferences with uncertain categorization. A crucial feature of this model is the conditional independence assumption—it assumes that the within category feature correlation is zero. In prior research, this model has been found to provide a poor fit to participants' inferences. This evidence is restricted to task environments inconsistent with the conditional independence assumption. Currently available evidence thus provides little information about how this model would fit participants' inferences in a setting with conditional independence. In four experiments based on a novel paradigm and one experiment based on an existing paradigm, we assess the performance of Anderson's model under conditional independence. We find that this model predicts participants' inferences better than competing models. One model assumes that inferences are based on just the most likely

category. The second model is insensitive to categories but sensitive to overall feature correlation. The performance of Anderson's model is evidence that inferences were influenced not only by the more likely category but also by the other candidate category. Our findings suggest that a version of Anderson's model which relaxes the conditional independence assumption will likely perform well in environments characterized by within-category feature correlation.

Keywords Feature inferences · Categories · Predictions · Judgments · Rational analysis · Bayesian model

According to J. Anderson, 'The basic goal of categorization is to predict the probability of various inexperienced features of objects' (Anderson, 1991). At the same time, humans often find themselves in situations where the categories of the objects they perceive are uncertain. How do people make predictions about unobserved features of an object when the category of that object is uncertain?

A highly influential answer to this question is J. Anderson's 'rational model' (Anderson, 1991). Consider a setting where an individual observes a feature of an object and makes a prediction about an unobserved feature of that object. The values of the two features are denoted by X (first feature) and Y (second feature). It is assumed that the individual has organized her knowledge of the domain in a set of categories $\mathcal{C}$.¹ According to Anderson's model (AM), the probability that the value of the second feature is y when

---

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13423-017-1372-y>) contains supplementary material, which is available to authorized users.

✉ Elizaveta Konovalova
elizaveta.konovalova@upf.edu

Gaël Le Mens
gael.le-mens@upf.edu

¹ Barcelona Graduate School of Economics, Barcelona School of Management, Universitat Pompeu Fabra, Barcelona, Spain

² University of Southern Denmark, Odense, Denmark

¹When we refer to 'Anderson's model' we only refer to the feature inference component of the original model. The original model had additional components to allow it to learn categories. This way of referring to Anderson's model is similar to Murphy and Ross (2010a).

the individual knows that the value of the first feature is x is given by

$$P(y | x) = \sum_{c \in \mathcal{C}} P(c | x) P(y | c), \quad (1)$$

where $P(c | x)$ is the subjective probability that the object comes from category c given the observed feature value x and $P(y | c)$ is probability that the second feature has value y given that the object belongs to category c . An important qualitative prediction of this model is that people take into account all the candidate categories when making an inference about the unobserved feature Y on the basis of the value of the observed feature $X = x$.

A large amount of empirical work has focused on testing this prediction. Existing findings are mixed. Some experimental evidence suggests that participants' inferences are the same as those implied by a model that relies only on the most likely category given the observed feature (the 'target category') (Chen, Ross, & Murphy, 2014a, b; Malt, Ross, & Murphy, 1995; Murphy & Ross, 1994, 2010a; Murphy, Chen, & Ross, 2012; Ross & Murphy, 1996; Verde, Murphy, & Ross, 2005). Other experiments suggest that participants rely on more than just the target category (Chen et al., 2014a; Hayes & Chen, 2008; Hayes & Newell, 2009; Murphy & Ross, 2010a; Newell, Paton, Hayes, & Griffiths, 2010; Verde et al., 2005). Finally, still other experiments suggest that participants do not pay attention to categories at all but instead are sensitive to the overall feature correlation (Griffiths, Hayes, Newell, & Papadopoulos, 2011; Griffiths, Hayes, & Newell, 2012; Hayes, Ruthven, & Newell, 2007; Papadopoulos, Hayes, & Newell, 2011). Several recent papers have attempted to uncover the conditions under which people are more likely to rely on multiple categories or just the target categories. For example, Murphy and Ross (2010b) found that participants were more likely to use multiple categories when the most likely category gives an ambiguous inference, and less likely to do so when the most likely category gives an unambiguous inference. Chen et al. (2014a) found that participants' inferences were likely to be influenced by multiple categories when the inference was implicit, whereas they were likely to be influenced by just the target category when the inference was explicit. Griffiths et al. (2012) found that participants' inferences were more likely to be influenced by a single category when participants had been trained to classify stimuli before the feature induction task.

Despite the diversity of findings, the studies that analyzed the performance of Anderson's model converge in showing that it provides a poor fit to experimental data. Central to this model is an assumption about the structure of the environment: it assumes that the within-category feature

correlations are equal to 0 (this is the '*conditional independence*' assumption). We believe this model can be seen as a 'rational model' only to the extent that this assumption is consistent with the structure of the actual task environment. We reviewed all prior experiments on feature inference with uncertain categorization (reported in the papers cited above) to check whether the task environments of these experiments were characterized by conditional independence. We found that it is the case in *none* of the previously published experiments.²

The poor performance of Anderson's model in an environment without conditional independence suggests that people do not make this assumption in such environments (a point made by (Murphy & Ross, 2010a)). Yet, currently available evidence provides little information about how this model would fit participants' inferences in a setting where conditional independence is satisfied. How well would Anderson's model (AM) predict participants' inferences in a task environment consistent with the conditional independence assumption?

At first sight, this question might seem moot. After all, Murphy and Ross (2010a) noted that there are many environments in which this assumption is not satisfied. For example, they argued that within-category feature correlation can result from large category difference. One example is sexual dimorphism in animals (Murphy & Ross, 2010a, p. 14). Male deer are larger and have different coloration than females. Therefore, these features are correlated within the category 'deer'. Similar feature correlations are present in consumer goods categories like books or computers. There is also evidence that people are aware of some within-category correlations (Malt & Smith, 1984).

However, even if there are possibly few naturally occurring environments that satisfy conditional independence, it is important to assess the performance of the Anderson's model in such settings. This is because there currently does not exist a rational model for environments where the conditional independence assumption does not hold. If Anderson's model performs well under conditional

²In Hayes and Newell (2009), the authors write that 'all the experimental categories were designed so that their component feature dimensions were statistically independent within and between categories' (p.733). Computations based on the statistical structure reported in their Table 1 show that this claim is not exact. Consider the Terragaxis category in the 'Divergent' condition. Here we have $P(\text{Rash} \& \text{Headache}) = 3/8$, $P(\text{Rash}) = 4/8$ and $P(\text{Headache}) = 7/8$. If Rash and Headache were independent within category, we would have $P(\text{Rash} \& \text{Headache}) = P(\text{Rash}) P(\text{Headache})$. This is clearly not the case. Therefore, the statistical structure of the experiments in this paper is not characterized by conditional independence. (This critique does not invalidate in any ways the results reported in that paper.)

independence—when it can be seen as a ‘rational model’—this will suggest that an extension of this model to settings without conditional independence needs to be developed. Such a model is likely to perform well.

We analyzed the performance of Anderson’s model in a task environment characterized by conditional independence, consistent with this key assumption of the model. In five experiments, we found that the model performed better than other competing models. This finding is important because it suggests that people’s inferences can be influenced by several categories when making inferences under uncertain categorization. Although there already exists some evidence that this can be the case (e.g., Chen et al., 2014a; Griffiths et al., 2012; Murphy & Ross, 2010b), we explain below that such evidence is based on a design that does not allow the parsing out between two possible interpretations of the data: that participants ignore categories altogether or that categories influence inferences in a fashion close to what would be predicted by application of Bayes’ theorem. The results reported in this paper suggest the later interpretation.

In the following, we describe the existing experimental paradigm that has been used by most of the literature on feature inference under uncertain categorization. We explain how the fact that it relies on discrete-valued features makes it of limited usefulness to the performance of assess Anderson’s model. Then we introduce our adaptation of Anderson’s model to continuous environments and describe competing models. Subsequently, we report the performance of Anderson’s model in four experiments based on a novel paradigm with continuous features and one experiment based on the existing paradigm with discrete features. Finally, we discuss how our findings relate to prior research.

Existing paradigm - discrete features

In the experimental paradigm used in the vast majority of experiments that focused on feature prediction with uncertain categorization, participants are shown a set of items of various shapes and colors divided into small number of categories, typically four (Murphy & Ross, 1994). Then they are told that the experimenter has a drawing of a particular shape and were asked to predict its likely color (or similar questions about the probability of an unobserved feature given an observed feature). An important characteristic of this paradigm is that the categories are shown graphically to the participants. The idea was to avoid complications related to memory and category learning by participants.

Suppose the two features are X and Y and there are four categories. Participants are asked to estimate $P(y | x)$, the proportion of items with $Y = y$ out of items with $X = x$.

There is some evidence that participants’ predictions are the same as those implied by a model that focuses on just the ‘target’ category, that is, the most likely category given the observed feature (Murphy & Ross, 1994). There is also some evidence that participants sometimes make predictions that are the same as those implied by a model that takes into account multiple categories (Murphy & Ross, 2010a). Still, other experiments have found evidence that participants do not pay attention to categories at all but instead are sensitive to the overall feature correlation (Hayes et al., 2007; Papadopoulos et al., 2011; Griffiths et al., 2012).

A limitation of this paradigm pertains to the fact that the features are discrete-valued. This implies that the predictions of a model that ignores categories altogether or makes optimal use of the categories are exactly the same. This is a consequence of the law of total probability. In this case, we have

$$P(y | x) = \sum_{c=1}^4 P(c | x)P(y | cx), \quad (2)$$

where $P(c | x)$ is the proportion of items belong to c out of all the items such that $X = x$, and $P(y | cx)$ is the proportion of items with $Y = y$ out of the items that both are in c and have $X = x$.

In settings where there is conditional independence, we have $P(y | cx) = P(y | c)$ and thus the above equation can be rewritten as:

$$P(y | x) = \sum_{c=1}^4 P(c | x)P(y | c). \quad (3)$$

In order to estimate $P(y | x)$, a participant that would ignore the categories would consider all objects with $X = x$ and would respond with the proportion of objects with y among all objects with x . A participant that would consider all four categories would compute the proportion of items with y among the items with x in each category and then would compute the weighted average by multiplying each of these numbers by her estimates of $P(c | x)$. The responses given by the two participants would be *exactly the same*. It is therefore difficult to assess whether the participants use multiple categories (but see Murphy & Ross, 2010a for an attempt to do so using post-prediction questions). When features are continuous, however, the predictions of these two strategies differ.

Below, we describe a version of Anderson’s model adapted to a continuous environment and report four experiments designed to test this model. We return to the discrete environment setup in Experiment 5 and the General Discussion section.

Rational feature inferences in a continuous environment

Representing mental categories

We depart from the prior literature on feature inference with uncertain categorization by focusing on a setting with continuously valued (as opposed to discrete) features. Following recent work, we model mental categories using probability distribution functions (pdfs) on the feature space (Ashby & Alfonso-Reese, 1995; Sanborn, Griffiths, & Shiffrin, 2010). Let $c \in \mathcal{C}$ be a category. We denote by $f(x, y | c)$ the value of the associated pdf at position (x, y) in the feature space, where x denotes the value of the first feature and y denotes the value of the second feature. This pdf denotes the prior belief of the individual over positions given that she knows that an object is from category c .

For simplicity, in what follows we assume there are two relevant categories ($\mathcal{C} = \{1, 2\}$) each represented by bi-variate normal distributions (Ashby & Alfonso-Reese, 1995):

$$\begin{pmatrix} X_c \\ Y_c \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_{xc} \\ \mu_{yc} \end{pmatrix}; \begin{pmatrix} \sigma_{xc}^2 & 0 \\ 0 & \sigma_{yc}^2 \end{pmatrix} \right), \quad (4)$$

where μ_{xc} and μ_{yc} are the category means for the two features, and σ_{xc} and σ_{yc} are the standard deviations. Consistent with the conditional independence assumption, the within-category feature correlation is zero. See Fig. 1 for an example.

Anderson's rational model (AM)

By adapting Eq. 1 to this continuous setting, we express the posterior on the second feature given the value of the first feature:

$$f(y | x) = \sum_{c \in \mathcal{C}} P(c | x) f(y | c), \quad (5)$$

where $P(c | x)$ is the subjective probability that the object comes from category c given that the first feature is observed to have value x and $f(y | c)$ is the marginal distribution of the second feature, conditional on the fact that the object is a c .

Anderson's model assumes that the subjective probabilities of the candidate category are given by Bayes' theorem:

$$P(c | x) = \frac{P(c) f(x | c)}{f(x)} = \frac{P(c) \int_v f(x, v | c) dv}{\sum_{c \in \mathcal{C}} P(c) \int_v f(x, v | c) dv}, \quad (6)$$

where $P(c)$ is the prior on the category.

In the special case with two categories and normally distributed category pdfs, we have:

$$f(y | x) = P(c_1 | x) f_{\mu_{y1}, \sigma_{y1}}(y) + P(c_2 | x) f_{\mu_{y2}, \sigma_{y2}}(y), \quad (7)$$

where f_{μ_y, σ_y} denotes the density of a normal distribution with mean μ_y and standard deviation σ_y , $P(c_2 | x) = 1 - P(c_1 | x)$, and

$$P(c_1 | x) = \frac{1}{1 + e^{ax^2 - bx + c}}, \quad (8)$$

with

$$a = \frac{\sigma_{x2}^2 - \sigma_{x1}^2}{2\sigma_{x2}^2\sigma_{x1}^2},$$

$$b = \frac{\sigma_{x2}^2\mu_{x1} - \sigma_{x1}^2\mu_{x2}}{\sigma_{x2}^2\sigma_{x1}^2},$$

$$c = \frac{\sigma_{x2}^2\mu_{x1}^2 - \sigma_{x1}^2\mu_{x2}^2}{2\sigma_{x2}^2\sigma_{x1}^2} + \log \frac{\sigma_{x2}}{\sigma_{x1}} + \log \frac{P(c_2)}{P(c_1)}.$$

We assume that the priors on the two categories, $P(c_1)$ and $P(c_2)$, are both equal to 0.5.

Fig. 1 Categories used in the four experiments. Participants were shown the level of ‘Rexin’ (x-axis) and were asked to predict the level of ‘Protropin’ (y-axis). The categories are ‘rat’ (R) and ‘mouse’ (M). $\mu_{xR} = 80$, $\mu_{yR} = 55$, $\mu_{xM} = 50$, $\mu_{yM} = 75$, $\sigma_{xR} = \sigma_{xM} = 10$, $\sigma_{yR} = \sigma_{yM} = 1$

Competing models

Prior literature suggests that people frequently focus on the most likely category and that they sometimes ignore categories altogether but are sensitive to the overall feature correlation. We describe translations of these perspectives to the continuous environment.

Single category - independent features (SCI) We refer to the most likely category given the observed feature (x) as the ‘target’ category (this is category 1 if $P(c_1 | x) > .5$, as given by Eq. 8). The posterior has the same structure as in Anderson’s model, but with all the weight on the target category (c^*). In this case, $f(y | x) = f_{c^*}(y | x)$, where $f_{c^*} = f_{\mu_{y1}, \sigma_{y1}}$ if the target category is category 1, and $f_{c^*} = f_{\mu_{y2}, \sigma_{y2}}$ otherwise. The ‘switch’ is situated where x is such that $P(c_1 | x) = .5$. In the rest of the paper, we refer to this value as the ‘boundary’.

Linear model (LM) Prior literature considered the ‘feature conjunction’ approach as a model that is sensitive to the overall statistical association between the two features across objects, independently of categorical boundaries. This model simply computes the empirical probability of the unobserved feature given the observed feature based on all the data, ignoring categorical boundaries. A direct analogue in the continuous setting does not exist because the agent might have to infer Y conditional on an x value to which she has never been exposed. This observation implies that a model that ‘regularizes’ the available observations is in order. This could be a parametric model or a non-parametric exemplar model that weights prior observations based on their similarity to the stimulus (Ashby & Alfonso-Reese, 1995; Nosofsky, 1986). For the sake of simplicity, we analyze a linear model. This is the simplest model that takes into account the overall feature correlation:

$$f(y | x) = f_{a_0 + a_1 x, \sigma_l}(y), \quad (9)$$

where $f_{a_0 + a_1 x, \sigma_l}(y)$ denotes a normal *pdf* with mean $a_0 + a_1 x$ and standard deviation σ_l . The parameters are the coefficients of the best-fitting linear model based on the observed samples from the two categories.

Decision rule

The outputs of all three models, as described above, are posterior distributions: subjective probability distributions over the value of the second feature (Y) given the observed feature (X). To make empirical predictions about human inferences, we need to specify how this posterior distribution translates into responses. In analyses of our experimental results, we will assume that the response is a random draw from the posterior distribution—this is a ‘probability

matching’ decision rule. Other decision rules are theoretically possible. They would lead to different model predictions. We return to this issue in the General Discussion section of the paper.

Experiment 1

Participants faced a feature inference task that closely matches the setting of the previous section. Following standard practice in the study of feature inference with uncertain categorization, we used a ‘decision-only’ paradigm: participants were provided with a graphical depiction of the categories which remained visible when they made inferences about the second feature on the basis of the value of the first feature. We adopted this design to avoid issues related to memory.

Design

Our experiment used artificial categories to avoid the influence of domain-specific prior knowledge. We asked the participants to assume they were biochemists who studied the levels of two hormones in blood samples coming from two categories of animals (e.g., Kemp, Shafto, & Tenenbaum, 2012). The hormones were called ‘Rexin’ and ‘Protropin’ and the two categories of animals were ‘Mouse’ and ‘Rat.’ We provided the participants with visual representations of the categories in the form of scatter plots of exemplars of the two categories (see Fig. 1). In addition, participants went through a learning procedure designed to familiarize themselves with the position of the categories in feature space (see [Supplementary Material](#)). In the judgment stage, participants were asked to infer, without feedback, the likely level of Protropin, based on the level of Rexin, for 48 blood samples which didn’t indicate the animal they came from (the category was thus uncertain). The question was ‘What is the likely level of Protropin in this blood sample?’ Participants answered using a slider scale with minimal value 40, maximal value 90, and increments of 1 unit.

Thirty participants recruited via Amazon Mechanical Turk completed the experiment for a flat participation fee.³

Model predictions

Figure 2 depicts the posterior distributions, $f(y | x)$, implied by the three competing models. The posterior for AM is based on Eq. 7 and the Bayesian category weights given by Eq. 8. The posterior for SCI is based on Eq. 7 and the all-or-nothing category weights. The parameters are

³The stimuli and data from the five experiments are available on Open Science Framework: www.osf.io/wps39.

Fig. 2 Posteriors ($f(y | x)$) on the second feature (Rexin) given the value of the first feature (Protropin) of the competing models. The darker areas correspond to higher values of the posterior and lighter areas correspond to lower values. The vertical line at $x = 65$ represents the ‘boundary’, i.e., the value of x for which the two categories are equally likely $P(\text{‘rat’} | x = 65) = P(\text{‘mouse’} | x = 65)$

the coefficients used to generate the categories (see the legend of Fig. 1). The posterior for LM is based on Eq. 9. The parameters are the coefficients of the best fitting linear model based on the all the dots depicted on Fig. 1 (irrespective of their categories). With the stimuli used in the experiment, we have $a_0 = 94.8$, $a_1 = -0.45$ and $\sigma_l = 5.7$.

The crucial difference between the predictions of AM and SCI lies in the region around the x value at which both categories are equally likely (Rexin level of 65). Consider Rexin level of 60. According to SCI only high levels of Protropin (close to 75) are likely (the ones corresponding to the “Mouse” category). According to AM, however, both high (close to 75) and low levels (close 55) of Protropin are likely.

Results

Parameter-free model comparison Here we assume that the model parameters for AM and SCI are the coefficients used to generate the categories and that the parameters for LM are those of the best fitting regression line, just as in Fig. 2. We computed the log-likelihood fit of each model on a participant-by-participant basis.⁴ Anderson’s model (AM) is the best fitting model for the majority of participants (74% of them, see Table 1). Figure 3 shows the inferences of all

participants as well as the log-likelihood of the three models for each participant.

Comparison of models with parameters estimated participant-by-participant The comparison of the parameter-free models implicitly assumes that the participants perceived the categories accurately (i.e., the parameters of their category *pdfs* were exact). This might not have been the case, however. For example, participants might have misjudged the position of the point where categories are equally likely ($x = 65$). Inspection of Fig. 3 reveals that the perceived position of this ‘boundary’ is essential to the performance of the single category model (SCI). A slight error leads to a strong penalty in terms of log-likelihood that might not translate to the fact that a participant used multiple categories. For example, participant #4 made predictions that are clearly indicative of a focus on just one category since the predictions correspond to the median y -level for ‘Mouse’ (the category on the left) when x is low and to the median y -level for rats (the category on the right) when x is high. But the participant switched between categories not exactly at the ‘boundary’ of $x = 65$. This implies a strong penalty to the likelihood of the single category inference (SCI) model. A strict version of the SCI model discussed in the prior literature (e.g. Murphy & Ross, 1994) is thus a poor performer in our task. To give a better chance to the SCI model and account for possible misperception of the categories, we estimated the

⁴In all analyses, we removed two judgments where the given Rexin level was $x = 65$. At this value, the two categories are equally likely and the SCI does not make an explicit prediction.

Table 1 Percentage of participants whose feature predictions were best fit by each of the candidate models

Model	Exp. 1		Exp. 2		Exp. 3		Exp. 4	
	True	Est.	True	Est.	True	Est.	True	Est.
AM: Anderson	22(74%)	18(60%)	26(87%)	26(87%)	19(63%)	17(57%)	23(74%)	21(68%)
SCI: Single Cat.	4(13%)	12(40%)	1(3%)	3(10%)	0	9(30%)	3(10%)	8(26%)
Indep. Features								
LM: Linear	4(13%)	0	3(10%)	1(3%)	11(37%)	4(13%)	5(16%)	2(6%)
# participants	30		30		30		31	

parameters of each model on a participant-by-participant basis (by maximizing the likelihood).⁵

Table 2 reports the mean estimated parameter values (the mean was estimated across participants for whom the focal model is the best). The average parameter estimates are close to the true values for both the rational and the single category model. This suggests that participants used the categories we intended them to use. Models were compared in terms of the BIC criterion. For 60% of the participants, AM provides the best fit while SCI provides the best fit for the rest of the participants.

Analyses of the ‘switching’ behavior of the participants

A crucial prediction of Anderson’s rational model (AM) is that in the area where the ‘target’ category is uncertain (around the boundary at $x = 65$), there are oscillations between the typical level of Protropin (y -axis) for mice (about 75) and rats (about 55). Consider one participant in the experiment. Suppose the participant has to make an inference about a blood sample with a level of Rexin of 70. The probability that this sample is from a Rat is about 0.8. Anderson’s model predicts that in 80% of the cases, a participant facing this situation will give a response close to 55 (typical Protropin level for a Rat sample) and that in about 20% of the cases, she will give a response close to 75 (typical Protropin level for a Mouse sample). If we collect many such judgments in the area where the ‘target’ category is uncertain, we should expect that some inference values will be close to 55 and others close to 75. The top row of Fig. 4 shows the inferences of ten simulated participants who follow Anderson’s model. All these simulated participants show oscillations between Protropin levels around 55 and 75 (the x values used for the simulation are the same as those used in the experiments, without any instance of $x = 65$).

By contrast, no such oscillation is implied by the single category model (SCI). In this case, there is just one

‘switch’ at the boundary ($x = 65$). The bottom row of Fig. 4 shows the inferences of ten simulated participants who are assumed to follow the single category model. Inferences are close to 75 for Rexin levels lower than 65 and close to 55 for Rexin levels higher than 65.

Instances of the two distinct inference patterns can clearly be seen on the graphs depicting the inferences of the participants in the experiment (Fig. 3). For example, participant 4 switched exactly once at the ‘boundary’ whereas participant 8 switched many times between the two modal Protropin levels of 55 and 75. Our participant-by-participant model estimations identified this difference since the single-category model provides the best fit to the inferences of Participant 4 whereas Anderson’s model provides the best fit to the inferences of Participant 8.

In Experiment 1 there are 12 (40%) participants with exactly one switch. This is very close to the number of participants best fit by the single category model (with parameters estimated participant-by-participant – see Table 1).

Discussion

Most participants’ inferences are better explained by Anderson’s rational model (AM) than by the single category model (SCI) and the linear model (LM). In this experiment, we provided participants with a visual representation of the categories. One might wonder if this design captures the psychological process that underlies inferences with uncertain categorization when such graphical representation is not available at the time of the inference. It could be that participants engaged in some elaborate form of curve fitting on the basis of the graphs we showed to them. We address this potential concern in Experiments 2 & 3.

Experiments 2 & 3

The experiments follow a design similar to Experiment 1. We recruited 30 participants via Amazon Mechanical Turk for each experiment.

⁵We used box-constraint optimization with six free parameters for AM and SCI. The lower bound for all parameters was 0. LM was optimized over three parameters without constraints.

Fig. 3 Inference of the participants of Experiment 1. The log-likelihoods of each parameter-free model is shown on each participant's graph. *AM* Anderson's rational model, *SCI* single category independent feature model, *LM* linear model. The red font indicates the best-fitting model

Design of Experiment 2 The only change in comparison to Experiment 1 is that we removed the graphical representation of the categories (i.e., the graph of Fig. 1) on the screens on which participants made judgments (during learning and test stages). The graph was shown in the instructions and before every judgment, but not on the judgment screen.

Design of Experiment 3 In this experiment, participants never saw any graphical representation of the data. They learnt the categories from experience by first seeing 40 exemplars of both categories (Rexin and Protropin values), then making within-category inferences (of Protropin level based on Rexin level) with feedback and categorizations of

blood samples as Rat or Mouse, based on Rexin level (see [Supplementary Material](#)).

Results Experiments 2 & 3

Parameter-free model comparison Removing the graphical depiction of the data didn't drastically change the pattern of results. Just as in Experiment 1, Anderson's model is the best for the majority of the participants in both experiments (Table 1).

Comparison of models with parameters estimated participant-by-participant and switching behavior In comparisons based on the BIC, Anderson's model (*AM*) is

Table 2 Estimated model parameters

	Parameter	$\mu_{x,R}$	$\mu_{y,R}$	$\mu_{x,M}$	$\mu_{y,M}$	σ_x	σ_y	a_0	a_1	σ_l
	True Value	80	55	50	75	10	1	94.8	-0.45	5.7
Experiment	Model									
1	<i>AM</i>	81.4	55.2	44.9	74.9	11.5	1.0			
	<i>SCI</i>	82.0	55.3	45.9	74.7	12.9	1.0			
	<i>LM</i>							88.3	-0.4	6.6
2	<i>AM</i>	75.7	55.4	44.9	75.3	10.7	0.9			
	<i>SCI</i>	81.6	55.1	51.9	75.2	15.5	0.4			
	<i>LM</i>							24.1	0.6	3.9
3	<i>AM</i>	77.9	54.9	43.2	74.9	11.6	1.0			
	<i>SCI</i>	79.3	55.7	49.9	73.8	11.4	2.9			
	<i>LM</i>							70.6	-0.1	6.0
4	<i>AM</i>	81.7	55	48.2	75.2	17.5	0.9			
	<i>SCI</i>	82.7	55	44.9	74.9	24.9	0.7			
	<i>LM</i>							68.8	0.1	4.9

Parameters were estimated separately for each participant. The values are the mean estimated parameters across participants for whom that model is the best. *AM* Anderson's rational model, *SCI* single category independent feature model, *LM* linear model

Fig. 4 Predictions of AM and SCI of inferences of ten artificial participants. The top row shows predictions of AM, the bottom row shows the predictions of SCI

by far the best fitting model (Table 2). As in Experiment 1, the single-category model (SCI) performs better in this comparison than in the comparison of parameter-free models, but worse than Anderson's model.

In Experiments 2 & 3 there are 4 (13%) and 8 (28%) participants with exactly one switch (see also participant-by-participant inferences in Figures S3 & S4 in the Supplementary Material). These numbers closely reflect the performance of the single-category model (with parameters estimated participant-by-participant).

Discussion of Experiments 1–3

Taken together, Exp. 1–3 show that Anderson's model (AM) provides a better fit to the data than the single category model (SCI). The linear model provides a very poor fit to the data. This pattern of results is consistent across experiments, in which we varied the information available about the categories. Whether participants saw a graphical representation of the categories in feature space at time of inference (Exp. 1), this representation was seen in the learning stage but removed in the inference stage (Exp. 2), or never seen (Exp. 3), the patterns of feature inferences were similar.

We would like to claim that good performance of Anderson's model is evidence for the integration of information across categories. In other words, we would like to claim that people use a cognitive algorithm of the following kind:

1. observe $X = x$;
2. compute the posterior distribution $f(y | x)$ according to Eq. 7;

3. provide an estimate of Y by generating a random draw from the posterior.

Because the posterior depends on the marginal distributions of the unobserved features for both the target and the non-target category, we call this cognitive algorithm 'AM non-target'.

The evidence gathered so far does not unequivocally show that participants used this kind of cognitive algorithm. The reason is that the results of Exps. 1–3 are also compatible with a noisy version of the single category inference (SCI) model. Suppose a participant uses SCI, but is uncertain about the location of the boundary at which the 'Rat' category becomes more likely than the 'Mouse' category. Let β denote the uncertain position of the boundary on the x -axis. Suppose inferences about y are produced by the following cognitive algorithm:

1. observe $X = x$ and estimate the position of the boundary β ;
2. evaluate if $x < \beta$ or if $x > \beta$.
3. if $x < \beta$, select the 'Mouse' category; else select the 'Rat' category. Denote the selected category by c^* .
4. provide an estimate of Y given $X = x$ by producing an intuitive estimate of the mode of the posterior distribution conditional on the selected category $f(y | c^*)$.

Assume, moreover, that the uncertainty is such that the participant's belief about the location of this boundary is represented by a probability density $g(\beta) = \frac{\partial P(Rat|\beta)}{\partial \beta}$. In this case, the inferences produced by this algorithm are compatible with Anderson's rational model: $P(Rat | \beta)$ follows Eq. 8, assuming $c_1 = Rat$ and $x = \beta$ —see [Supplementary Material](#) for an explicit formulation of $g(\beta)$. Figure 5

Fig. 5 Density of the position of the uncertain boundary between the categories ‘Rat’ and ‘Mouse’ implied by the second cognitive algorithm (SCI with uncertain boundary)

depicts the density of the uncertain boundary, $g(\beta)$. It is a unimodal symmetric distribution centered at the mid-point between the two categories ($x = 65$). We will refer to this algorithm as ‘SCI with uncertain boundary’.

If participants whose inferences are best fit by Anderson’s model rely on the ‘SCI with uncertain boundary’ cognitive algorithm, then eliminating the uncertainty about the boundary should reduce the fit of Anderson’s model as compared to SCI. This should not happen if people integrate information from both categories in inferring the unobserved feature value—if they use the ‘Anderson non-target’ algorithm. We designed an experiment that relied on these predictions.

Experiment 4

Experiment 4 used a design identical to Exp. 1, with one change: we provided participants with information that ruled out subjective uncertainty about the boundary at which one category becomes more likely than the other. Consider the following two hypotheses:

- H1: People whose inferences are best fit by Anderson’s rational model use the ‘SCI with uncertain boundary’ cognitive algorithm.
- H2: People whose inferences are best fit by Anderson’s rational model use the ‘AM non-target’ cognitive algorithm.

If hypothesis H1 is true, then removing the subjective uncertainty about the boundary should lead to inferences consistent with the SCI model. Therefore, under this

hypothesis, the SCI model should provide a better fit to participants’ inferences than in Exp. 1. If hypothesis H2 is true, removing subjective uncertainty about the boundary should not lead to a relative increase in the performance of the SCI model.

Design

Thirty-one participants recruited via Amazon Mechanical Turk completed the experiment. The design was identical to Experiment 1 except for the addition of the following note below the graph depicting the blood sample data: “Note: A blood sample with a Rexin level equal to 65 is equally likely to come from a Rat or a Mouse.” (Fig. S2 in Supplementary Material). The note was shown whenever the graph was shown. We checked that the participants understood the meaning of the note about the boundary in three true-false questions (see Supplementary Material).

Results

Comparison parameter-free models Anderson’s model is the best fitting model for 74% of the participants (see also Fig. S5 in Supplementary Material). This performance is very similar to that of Exps. 1–3. The performance of Anderson’s model is even higher among the 24 participants who passed the comprehension check (AM provides the best fit for 88% of these participants).

Comparison of models with parameters estimated participant-by-participant and switching behavior Maximum likelihood estimation of the parameters and the

Fig. 6 One of the panels used in Experiment 5

analysis of the switching behavior yields results similar to those of Exps. 1–3 (Table 2). There are eight out of 31 participants (26%) with exactly one switch.

Discussion of Experiment 4

Informing the participants of which category was the more likely did not lead to a performance improvement of the single category model (SCI) model relative to Anderson’s model (AM). This is inconsistent with Hypothesis H1 but consistent with Hypothesis H2. This suggests that participants who made inferences consistent with Anderson’s rational model (in Exps. 1–3) did not use the ‘SCI with uncertain boundary’ cognitive algorithm. Rather, it likely reflects the operation of a cognitive algorithm in which participants’ inferences were affected not only by the more likely category but also by the non-target category.

Discussion of Experiments 1–4

The first four experiments analyzed the performance of Anderson’s rational model in a task environment characterized by conditional independence. We found that in these four experiments, Anderson’s model performed well—unequivocally better than the single category inference (SCI) model. In other words, participants’ inferences were influenced by the non-target category. This seems to contradict the findings of prior studies that found evidence in favor of the SCI model.

A skeptic might wonder whether the difference between our results and the results of these studies could be due to the fact that our design differs in many ways from the standard paradigm used to study feature inference under uncertain categorization.⁶ After all, our design differs from this paradigm not only in terms of satisfying the conditional independence assumption, but also on other dimensions:

continuously valued features versus discrete features, flow of the experiment, number of categories, etc. We believe that relying on a discrete paradigm to study the performance of Anderson’s model is suboptimal because in this context the predictions of Anderson’s model and the predictions of the feature conjunction model (a model that ignores categories altogether) are the same, as explained in the first section of the paper. Nevertheless, we wanted to assess the performance of Anderson’s model in a setup that matched the standard paradigm as closely as possible. This is the purpose of the next experiment.

Experiment 5

The experiment was designed to fit within the standard discrete feature paradigm while satisfying the conditional independence assumption.

Design The design of this study closely follows the designs of Experiment 1 and 2 in Murphy and Ross (2010b) and Experiment 1 in Murphy and Ross (2010a). Just as in the original experiments, participants were shown drawings by four children (see Fig. 6 for one of the panels used in the experiment). Each category consisted of nine drawings and each drawing was a colored shape. The stimuli were designed such that in each category they were conditionally independent. For example, in the ‘Kyle’ category, there were four orange circles, two green circles, two orange squares, and one green square. We have $P(\text{orange} \& \text{circle}) = 4/9$. We also have $P(\text{orange})P(\text{circle}) = 6/9 * 6/9 = 4/9$.

The flow of the experiment was as follows. First, participants were shown general instructions about the children drawings. They were told each collection of nine children drawings was a sample from a larger set of drawings by this child. Each question had several parts, as shown in the following example:

- I have a new drawing that is a circle.
- What color do you think this circle drawing has?
- What is the probability (0–100%) that it is has this color?

⁶We thank an anonymous reviewer for expressing this skepticism.

The instructions also explained that 0 on the scale means that it is impossible, 50% means that this happens about half of the time, 100% means that this is completely certain. Participants answered four such questions, two about one panel, and two about another similar panel. For each panel, one question used shape as a given feature and the other one used color.

In most of the original studies (e.g., Experiment 1 in Murphy & Ross, 2010b), participants were asked about the target category after being informed of the feature (e.g., that the shape was a circle) and before making their prediction (e.g., about the color of the circle). In the above example, participants would be asked “Which child do you think drew it?” after reading “I have a new drawing that is a circle”. They were also asked for their confidence in that judgment (“What is the probability (0–100%) that the child you just named drew this?”). Because prior research suggests that this question makes people more likely to rely on the SCI strategy (e.g., Murphy & Ross, 2010b), we created three between-subject conditions. In the ‘NO’ condition, participants were not asked about the target category. In the ‘PR’ condition, participants used four sliders to indicate the probabilities that each child had drawn the shape (the sum of the four probabilities was constrained to be 100). In the ‘MC’ condition, participants answered a multiple-choice question (with the four children as possible answers) about the target category and their confidence in that judgment, as described at the beginning of this paragraph.

In order to closely match the design of prior experiments, there were two within-subject conditions. In the *agree* condition, the predictions of the SCI and the AM models were

the same. In the *disagree* condition, the predictions of the SCI and AM models differed. For example, with the panel of Fig. 6 the question in the *agree* condition was about the shape of a green drawing. In this case, both models predict that the response is ‘square’. The question in the *disagree* condition was about the color of a ‘circle’. Here SCI predicts ‘orange’ (with ‘Kyle’ as the target category) whereas AM predicts ‘green’. See Table 3 for the details on the models’ predictions.

Ninety-two participants recruited via Amazon Mechanical Turk completed the experiment for a flat participation fee. There were 31 participants in the NO condition, 30 participants in the PR condition, and 31 participants in the MC condition.

Results Unsurprisingly, the majority of participants’ responses for the *agree* condition questions were consistent with the predictions of AM and SCI (which are the same). The proportion of predicted responses were 61, 67, and 76% in the NO, PR, and MC conditions, respectively. (Each participant answered one question in both *agree* and *disagree* conditions for each panel; the resulting numbers of observations are thus 62, 60, and 62 for NO, PR, and MC conditions, respectively.)

We now turn to the results of the *disagree* condition questions. For each condition we calculated the proportion of participants’ responses that correspond to the predictions of the AM and SCI models. In the NO and PR conditions, participants’ inferences were much more consistent with Anderson’s model than with the single-category model. The response predicted by AM was chosen 65 and 53% of time

Table 3 Predictions of the models for the color of a circle for panel on Fig. 6

Response	Model		
	AM	SCI	LM
P(orange circle)	$P(\text{Kyle} \text{circle}) \cdot P(\text{orange} \text{Kyle}) + P(\text{Hanna} \text{circle}) \cdot P(\text{orange} \text{Hanna}) + P(\text{Ch} \text{circle}) \cdot P(\text{orange} \text{Ch}) + P(\text{Liz} \text{circle}) \cdot P(\text{orange} \text{Liz}) = 6/12 \cdot 6/9 + 3/12 \cdot 0 + 0 \cdot 6/9 + 3/12 \cdot 0 = 4/12$	$P(\text{orange} \text{Kyle}) = 6/9$	$= 4/12$
P(green circle)	$P(\text{Kyle} \text{circle}) \cdot P(\text{green} \text{Kyle}) + P(\text{Hanna} \text{circle}) \cdot P(\text{green} \text{Hanna}) + P(\text{Ch} \text{circle}) \cdot P(\text{green} \text{Ch}) + P(\text{Liz} \text{circle}) \cdot P(\text{green} \text{Liz}) = 6/12 \cdot 3/9 + 3/12 \cdot 6/9 + 0 \cdot 3/9 + 3/12 \cdot 3/9 = 5/12$	$P(\text{green} \text{Kyle}) = 3/9$	$= 5/12$
P(purple circle)	$P(\text{Kyle} \text{circle}) \cdot P(\text{purple} \text{Kyle}) + P(\text{Hanna} \text{circle}) \cdot P(\text{purple} \text{Hanna}) + P(\text{Ch} \text{circle}) \cdot P(\text{purple} \text{Ch}) + P(\text{Liz} \text{circle}) \cdot P(\text{purple} \text{Liz}) = 6/12 \cdot 0 + 3/12 \cdot 3/9 + 0 \cdot 0 + 3/12 \cdot 0 = 1/12$	$P(\text{purple} \text{Kyle}) = 0$	$= 1/12$
P(blue circle)	$P(\text{Kyle} \text{circle}) \cdot P(\text{blue} \text{Kyle}) + P(\text{Hanna} \text{circle}) \cdot P(\text{blue} \text{Hanna}) + P(\text{Ch} \text{circle}) \cdot P(\text{blue} \text{Ch}) + P(\text{Liz} \text{C}) \cdot P(\text{blue} \text{Liz}) = 6/12 \cdot 0 + 3/12 \cdot 0 + 0 \cdot 0 + 3/12 \cdot 6/9 = 2/12$	$P(\text{blue} \text{Kyle}) = 0$	$= 2/12$

AM Anderson’s model, SCI single-category independent features model, LM linear model or feature conjunction model

Table 4 Results of Experiment 5

Model	Condition		
	NO	PR	MC
AM: Anderson	40 (65%)	32 (53%)	22 (35%)
SCI: Single Cat.	12 (19%)	18 (30%)	33 (53%)
Indep. Features			
χ^2	24.1	5.8	3.3
95CI	[0.28; 0.62]	[0.05; 0.42]	[−0.36; 0.01]
# observations	62	60	62
# participants	31	30	31

Percentage of participants who's feature inferences corresponded to the models' predictions. For proportion comparison value of Person's χ^2 test statistic and confidence interval of the proportion difference are reported

in the NO and PR conditions, respectively. By contrast, the response predicted by SCI was chosen much less frequently at 19 and 30% of the time, respectively. The differences in proportions are significant (see Table 4 for details).

The results differ in the MC condition. In this case, SCI provides a much better fit to participants' inferences. The response predicted by SCI was chosen 53% of the time, whereas the response predicted by AM was chosen 35% of the time.

We find that in the “PR” and “NO” conditions the answer that corresponds to AM prediction was chosen significantly more (by 53 and 65% of the participants respectively, see Table 4 for details). In the “MC” condition, an answer that corresponds to SCI prediction was chosen by 53% of the participants. This is marginally significantly (at the 10% level) higher than the proportion of answers that corresponds to AM prediction.

Discussion The purpose of this experiment was to evaluate the performance of Anderson's rational model in an environment with discrete features and conditional independence. We found that Anderson's model performs well provided that participants were not asked to categorize the shape before formulating their feature inference. This suggests that participants were able to take into account more than just the target category, as in Exps. 1 to 4. Whether they considered the various candidate categories and weighted them optimally (or close to optimally) or they ignored categories altogether cannot be decided on the basis of these data, because the prediction of Anderson's model (optimal weighting of the candidate categories) and of the feature conjunction strategy (ignoring the categories) are the same. Taken together with the results of Exps. 1 to 4,

however, the results of this experiment suggest that participants were sensitive to categories and were influenced by more than just the target category when formulating their feature inferences.

Although our focus here is not on how categorization questions affect inferences, it is worth discussing our results about the three between-subject conditions (NO, PR, and MC conditions). We found that when participants were asked to select the most likely category of the object (‘MC’ condition), the single category inference model was the best-fitting model. In the other conditions, Anderson's model was the best-fitting model. This suggests that the question format strongly affects the extent to which feature inferences are influenced by one or several categories. These are consistent with the results of the experiments reported by Murphy and Ross (2010b). In their experiments, they found that when participants were asked to evaluate the probabilities of the four candidate categories, their feature inferences were influenced not only by the target category but also by the other candidate categories (the design of their Experiment 1 is closest to our MC condition and the design of their Experiment 2 is closest to our PR condition).

Related findings were reported by Hayes & Newell (2009, Exp. 3) and Murphy & Ross (1994, Exps. 5&6). In these experiments, the authors compared feature inference between a setting in which participants had to categorize items before the feature prediction (this condition is closest to our MC condition) and a setting where they did not have to make such categorization decision (this condition is closest to our NO condition). Both of the earlier studies found that inferences were more likely to be influenced by several categories when they did *not* follow a categorization decision. This finding is similar to the comparison of our MC and NO conditions: in the NO condition, inferences were much more likely to be influenced by several categories than in the MC condition.

Taken together, the comparisons of inferences between the MC condition on one side and the PR and NO conditions on the other side are consistent with a conjecture according to which much of the past evidence for the effect of a single category on feature inference could be due to the ordering of the questions in most experiments (categorization first and feature inference second). A comparison of inferences in the NO and PR conditions possibly supports this conjecture. In the NO condition, feature inferences were similar to those in the PR condition. This suggests that participants who were *not made to focus on just one category* by a categorization question readily consider more than just the target category (at least in the setting of our experiment, where the target category was not immediately clear). This in turn suggests that people's default strategy when the category is uncertain

is to consider more than one category. This default tendency could be overridden by having people focus on just one category, but more research is clearly needed to evaluate this conjecture.

General discussion

Integration of information over categories

This paper contributes to the literature that addresses the extent to which several categories affect feature inference when categorization is uncertain. In our studies participants' inferences were affected by several categories, which suggests that they integrated information across these categories. But what does it mean that participants *integrate* information across categories?

Information integration can happen at two levels in the model: when computing the posterior distribution and at the level of the decision rule. In our rendition of Anderson's model, information integration occurs in the computation of the posterior distribution. To see this, note that the posterior is bimodal when the more likely category is uncertain. Importantly, in this model, there is no information integration at the level of the decision rule. We assumed that the inferred feature value was a random draw from the posterior distribution. It is possible to think of alternative models where information integration also operates at the level of the decision rule. Maybe the most straightforward choice for such a decision rule is the expected value of the posterior $E[Y | x]$, computed with respect to the posterior produced by Anderson's model. Under conditional independence this is:

$$\begin{aligned} E[Y | x] &= P(c_1 | x)\mu_{y1} + P(c_2 | x)\mu_{y2} \\ &= P(c_1 | x)(\mu_{y1} - \mu_{y2}) + \mu_{y2}. \end{aligned} \quad (10)$$

In this case, the inferred feature value is the weighted average of the means of the categories. This 'AM Averaging' model is sensitive to the overall feature correlation, like the linear model we estimated on our data (LM), but it is more sophisticated: the mean of the posterior looks like a logistic curve (see Fig. 1 in Konovalova & Le Mens 2016 for an example). Such response curve predicts that in the area of uncertainty (x around 65) subjects would give y values that lie in between the category means (μ_{y1} and μ_{y2}). More generally, response models that rely on some form of averaging of the category-specific inferences to produce an inference in the area where the category is uncertain will produce a unimodal conditional distribution of response. In other words, if $g(y | x)$ denotes the conditional density of responses obtained for a specific x values, a model

that relies on averaging will be such that $g(\cdot | x)$ is a unimodal distribution⁷. This prediction is inconsistent with the evidence we obtained in our four experiments. Examination of the participant-by-participant data clearly shows that responses are bimodal in the area where the category is uncertain (x close to 65—see Fig. 3 and Figs. S3–S5 in the Supplementary Material).

To illustrate this further, we assessed the performance of the 'AM Averaging' model on our data. We specified the 'AM Averaging' model with a normally distributed error term (to account for the likely dispersion of responses around the deterministic prediction produced by human error). This results in the following posterior:

$$f(y | x) = f_{E[Y|x], \sigma_{AMA}}(y), \quad (11)$$

where $f_{E[Y|x], \sigma_{AMA}}(y)$ denotes a normal *pdf* with a mean $E[Y | x]$ (given by Eq. 10) and standard deviation σ_{AMA} .⁸ In comparisons of the AM, SCI and 'AM Averaging' models, we found that the 'AM Averaging' model provides the best fit to almost exactly the same number of participants as the linear model in the analyses reported above (see Table S1 in the Supplementary Material). In all cases, the number of participants best fit by the 'AM Averaging' model is substantially lower than the number of participants best fit by Anderson's model (AM).

This analysis suggests that most participants integrated information from the candidate categories in computing the posterior, but not in their decision rule. More work is clearly needed to understand which type of decision rule best explains feature inferences in the kind of continuous environment we have used in Exps. 1 to 4. Other decision rules are possible. For example, one could think of a decision rule that selects the mode of the posterior distribution, or a decision rule that consists of a random draw from a distribution that is centered around the mode. A possible way to study decision rules is to specify an explicit reward function. The form of the reward function will likely affect the decision rule used to formulate the inferences. A related question pertains to how the task environment affects the decision rule when no reward function is explicitly specified.

⁷Unless one assumes a very bizarre error term distribution.

⁸In the parameter-free version of the model, we took $\sigma_{AMA} = 4.7$. This value was obtained by maximum likelihood estimation of the model (11) with just σ_{AMA} as a free parameter and the true values for the other parameters ($\mu_{xR} = 80$, $\mu_{yR} = 55$, $\mu_{xM} = 50$, $\mu_{yM} = 75$, $\sigma_{xR} = \sigma_{xM} = 10$, $\sigma_{yR} = \sigma_{yM} = 1$), based on all the exemplars depicted on Fig. 1 (irrespective of their categories). In the version of the model with parameters estimated on a participant-by-participant basis, σ_{AMA} and all the other parameters were estimated by maximum likelihood, just as in the analyses of the experiments reported in earlier sections.

Relation to Nosofsky's exemplar model

In a recent paper, Nosofsky (2015) proposed an exemplar model that provides a good fit to existing data based on the discrete paradigm just discussed. Like ours, this model makes feature inferences that are influenced by all the categories. And for most parameter values, the target category receives a higher weight. The model relies on an assessment of the similarity between the observed stimulus and the data stored in memory and then does some similarity weighted prediction. The way the similarity is computed gives more weight to the most likely category (the 'target' category). It also gives some weight to the other categories. Therefore, just like our model, the exemplar model makes inferences that are influenced by several categories. This model differs from ours because the exemplar model is an algorithmic model whereas our model is specified at the computational level. Ours does not specify the details of the mental computations whereas the exemplar model does.

To cast light on the relation between Nosofsky's model and the other models discussed in this paper, we computed the predictions of Nosofsky's model in the context of the Experiment 5 reported earlier. The model has a parameter S , that characterizes the weight of exemplars that do not have the observed feature. The second parameter, L , regulates the sensitivity to the target category. Intuitively, when $L = 0$ the individual pays no attention to exemplars outside the target category. Let $\hat{P}(y | x)$ denote the probability that the second feature is equal to y if the first feature is observed to be equal to x according to the exemplar model (we use the 'hat' to emphasize that this is the model prediction). When $L = 0$, the model predicts:

$$\begin{aligned}\hat{P}(\text{orange} | \text{circle}) &= \frac{6}{9} \\ \hat{P}(\text{green} | \text{circle}) &= \frac{3}{9} \\ \hat{P}(\text{purple} | \text{circle}) &= 0 \\ \hat{P}(\text{blue} | \text{circle}) &= 0\end{aligned}$$

These are the same predictions as SCI (see Table 3). When $L = 1$ and $S = 0$, the exemplar model predicts

$$\begin{aligned}\hat{P}(\text{orange} | \text{circle}) &= \frac{4}{12} \\ \hat{P}(\text{green} | \text{circle}) &= \frac{5}{12} \\ \hat{P}(\text{purple} | \text{circle}) &= \frac{1}{12} \\ \hat{P}(\text{blue} | \text{circle}) &= \frac{2}{12}\end{aligned}$$

These are the same predictions as AM (see Table 3).

Future research should go beyond this specific case and clarify the exact links between the exemplar model and other computational models of feature inferences both in discrete and continuous environments. Shi, Griffiths, Feldman, and Sanborn (2010) have shown that exemplar models can be seen as algorithms for performing Bayesian inference provided the decision rule can be specified as the expectation of a function. This suggests that it might be possible to show that an exemplar model could approximate the prediction of the Anderson Averaging model discussed in the prior subsection (note that this model does not fit our experimental data well, however). It might be harder to design an exemplar model that can produce the same predictions as Anderson's model with a decision rule that consists in a random draw from the posterior. This is because this decision rule cannot easily be specified as the expectation of a function.

Conclusions

Taken together, our results show that in an environment characterized by conditional independence, Anderson's rational model is a good predictor of participants' inferences – a performance much higher than what was suggested by previous empirical research on inference with uncertain categorization. This good performance suggests that most participants were influenced not only by the most likely category (given the observed feature) but also by the other candidate category. There was heterogeneity among participants: a non-trivial minority of participants was best fit by the single-category model. This aspect of our results is consistent with prior research, which found heterogeneity in the propensity to rely on just one category—even though this research often found that most participants relied on the whereas we found relied on multiple categories (Murphy and Ross, 2010a).

Finally, it is important to note that our results do not speak of the realism of the conditional independence assumption. As explained in the introduction, there are many environments where conditional independence does not hold. We are currently adapting Anderson's model to environments with positive or negative within-category correlations. The results reported here suggest that such a model will perform well. Preliminary evidence suggests that it is the case.

Acknowledgements This paper is based on research reported in the Proceedings of the 38th Annual Conference of the Cognitive Science Society (Konovalova & Le Mens, 2016) although the data reported here is new. We are grateful for discussions with and comments by Mike Hannan, Robin Hogarth, participants in the Behavioral and Management Research Breakfast at UPF, the ConCats seminar at NYU, the 2015 Summer Institute at the MPI for Human Development, the SPUDM 2015 conference, and the JDM 2015 meeting. This

research was funded by the Spanish Ministry of Economics, Industry and Competitiveness Grants #PSI2013-41909-P and #AEI/FEDER UE-PSI2016-75353, Ramon-y-Cajal Fellowship #RYC-2014-15035 to Gaël Le Mens, Grant #BES-2014-068047 to Elizaveta Konovalova, the Spanish Ministry of Education Culture and Sports Jose Castillejo Fellowship #CAS15/00225 and BBVA Foundation grant #IN[15]_EFG_ECO.2281 to Gaël Le Mens. Gaël Le Mens thanks the University of Southern Denmark for its support. Emails: elizaveta.konovalova@upf.edu & gael.le-mens@upf.edu. The data from the 5 experiments are available on Open Science Framework: www.osf.io/wps39.

References

- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3), 409–429.
- Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39(2), 216–233.
- Chen, S. Y., Ross, B. H., & Murphy, G. L. (2014a). Decision making under uncertain categorization. *Frontiers in psychology*.
- Chen, S. Y., Ross, B. H., & Murphy, G. L. (2014b). Implicit and explicit processes in category-based induction: is induction best when we don't think? *Journal of Experimental Psychology: General*, 143(1), 227.
- Chen, S.Y., Ross, B.H., & Murphy, G.L. (2014). Decision making under uncertain categorization. *Frontiers in Psychology*, 5.
- Griffiths, O., Hayes, B. K., Newell, B. R., & Papadopoulos, C. (2011). Where to look first for an explanation of induction with uncertain categories. *Psychonomic Bulletin & Review*, 18(6), 1212–1221.
- Griffiths, O., Hayes, B. K., & Newell, B. R. (2012). Feature-based versus category-based induction with uncertain categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(3), 576–595.
- Hayes, B. K., & Newell, B. R. (2009). Induction with uncertain categories: When do people consider the category alternatives? *Memory & Cognition*, 37(6), 730–743.
- Hayes, B. K., Ruthven, C., & Newell, B. R. (2007). Inferring properties when categorization is uncertain: A feature-conjunction account. In *Proceedings of the 29th annual conference of the Cognitive Science Society* (pp. 209–214).
- Hayes, B. K., & Chen, T. H. J. (2008). Clinical expertise and reasoning with uncertain categories. *Psychonomic Bulletin & Review*, 15(5), 1002–7. <https://doi.org/10.3758/PBR.15.5.1002>
- Kemp, C., Shafto, P., & Tenenbaum, J. B. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology*, 64(1), 35–73.
- Konovalova, E., & Le Mens, G. (2016). Predictions with uncertain categorization: a rational model. In Trueswell, J., Papafragou, A., Grodner, D., & Mirman, D. (Eds.) *Proceedings of the 38th annual conference of the Cognitive Science Society* (pp. 722–727). Austin.
- Malt, B. C., & Smith, E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, 23(2), 250–269.
- Malt, B. C., Ross, B. H., & Murphy, G. L. (1995). Predicting features for members of natural categories when categorization is uncertain. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(3), 646–661.
- Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, 27(2), 148–193.
- Murphy, G. L., & Ross, B. H. (2010a). Category vs. object knowledge in category-based induction. *Journal of Memory and Language*, 63(1), 1–17.
- Murphy, G. L., & Ross, B. H. (2010b). Uncertainty in category-based induction: When do people integrate across categories?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(2), 263–276.
- Murphy, G. L., Chen, S. Y., & Ross, B. H. (2012). Reasoning with uncertain categories. *Thinking & Reasoning*, 18(1), 81–117.
- Newell, B. R., Paton, H., Hayes, B. K., & Griffiths, O. (2010). Speeded induction under uncertainty: The influence of multiple categories and feature conjunctions. *Psychonomic Bulletin and Review*, 17(6), 869–874.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1), 39.
- Nosofsky, R. M. (2015). An exemplar-model account of feature inference from uncertain categorizations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(6), 1929–1941.
- Papadopoulos, C., Hayes, B. K., & Newell, B. R. (2011). Noncategorical approaches to feature prediction with uncertain categories. *Memory & Cognition*, 39(2), 304–318.
- Ross, B. H., & Murphy, G. L. (1996). Category-based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(3), 736–753.
- Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, 60(2), 63–106.
- Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, 17(4), 443–464.
- Verde, M. F., Murphy, G. L., & Ross, B. H. (2005). Influence of multiple categories on the prediction of unknown properties. *Memory and Cognition*, 33(3), 479–487.