Informed inferences of unknown feature values in categorization

Michael J. Wood · Mark R. Blair

Published online: 18 November 2010
© Psychonomic Society, Inc. 2010

Abstract Many current computational models of object categorization either include no explicit provisions for dealing with incomplete stimulus information (e.g. Kruschke, Psychological Review 99:22–44, 1992) or take approaches that are at odds with evidence from other fields (e.g. Verguts, Ameel, & Storms, Memory &amp; Cognition 32:379–389, 2004). In two experiments centered around the inverse base-rate effect, we demonstrate that people not only make highly informed inferences about the values of unknown features, but also subsequently use the inferred values to come to a categorization decision. The inferences appear to be based on immediately available information about the particular stimulus under consideration, as well as on higher-level inferences about the stimulus class as a whole. Implications for future modeling efforts are discussed.

Keywords Category learning · Categorization · Inference · Missing information

It is a truism that humans must rely on incomplete information about the world in order to make decisions. When voting in an election, we are able to come to a decision in spite of the fact that we do not have exhaustive knowledge of the candidates' positions, personalities, and histories. Likewise, on a more basic level, we can recognize objects in the world without full information – for instance, we have all had the experience of recognizing someone from afar despite only seeing the back of their head.

Though it seems self-evident that we are able to categorize in the absence of perfect information, the ability to do so is not broadly reflected in the architecture of computational models of categorization. Models such as ALCOVE (Kruschke, 1992), EXIT (Kruschke, 2001), RASHNL (Kruschke & Johansen, 1999), and RULEX (Nosofsky, Palmeri & McKinley, 1994) explicitly or implicitly assume that a full complement of inputs will be provided – that information about every relevant stimulus feature is available for consideration in the categorization process. Verguts et al. (2004) pointed out the problems inherent in this approach, particularly in models involving geometric distance computations of similarity among exemplars. To make matters worse, missing input data can interact in unpredictable ways with the peculiarities of individual models' architecture to produce entirely unexpected and unreasonable predictions. Convincing the models to function within reasonable boundaries in the presence of missing data can be problematic, as demonstrated in the challenges Blair and Homa (2005) encountered when implementing the RASHNL model to fit a task with a variable number of stimulus dimensions.

While some models (e.g. SUSTAIN: Love, Medin & Gureckis, 2004) do include explicit provisions for dealing with missing data, the usual approach is to ignore any unknown stimulus features, factoring them out of similarity computations entirely. In their ADDCOVE model, Verguts et al. (2004) implemented an alternative approach in which similarity is computed on the basis of feature-matching, obviating the need for a special provision for missing data. Although these remedies constitute an improvement over models of categorization which make no provisions for situations in which only partial information is available, they are at odds with empirical work on how humans deal with missing data. Specifically, existing models assume an extremely low degree of flexibility among categorizers when working with incomplete information. The typical

M. J. Wood (✉) · M. R. Blair
Simon Fraser University,
Burnaby, British Columbia, Canada
e-mail: mw337@kent.ac.uk

approach is to treat all missing data in the same way, regardless of the context – by ignoring it, by according it a dummy value, and so on.

In this article, we seek to demonstrate that this one-size-fits-all approach is in error. Rather, the reality of the situation is quite complex: categorizers appear to make sophisticated inferences about the identity of missing stimulus features based on observed regularities among stored exemplars. Specifically, people are sensitive to both intercorrelations among features and higher-level inferences about the properties of the stimulus set.

The literature on missing information in categorization is quite sparse, but relevant research has been conducted in the fields of multivariate prediction and multiple-cue probability learning. The available literature suggests that rather than ignoring missing features, people instead infer default values and subsequently use those inferences in making decisions or predictions. The exact identity of the default value appears to change depending on the situation (Ganzach & Krantz, 1990; White & Koehler, 2004). In general, however, if a particular feature's value is unknown, people appear to infer the "mean" value, averaging over the values of that feature observed during previous experience.

This approach has a number of advantages. As observed by Ganzach and Krantz (1990), it results in the moderation of predictions from incomplete data, whereas inferring a more extreme value would result in a correspondingly extreme prediction – a strategy that would perhaps lead to maladaptive outcomes if widely applied. However, it is not immediately clear how broadly the tendency to infer the mean applies. Much of the previous research on missing information has been in the paradigm of multiple-cue probability learning rather than deterministic categorization, which has traditionally been the focus of models such as ALCOVE and SUSTAIN. While this is not necessarily a problem, it may be the case that an environment in which prediction is highly fallible and probabilistic encourages categorizers to hedge their bets when dealing with missing data, while more extreme inferences might be more common in a fully deterministic environment.

In addition, there has been relatively little examination of the effect of intercorrelated features on missing-feature inference. Covariation among real-world object features is very common, and indeed is one of the reasons why categorization is useful – for example, the presence of a spoiler on a car is generally associated with a relatively high-powered engine. Thus, it seems relevant to ask whether the tendency to infer mean values for unknown features is a general strategy in categorization, or if mean inference is simply the base case of a broader strategy of predicting unknown feature values on the basis of observed correlations with known features. White and Koehler (2004) put a significant number of potential strategies for missing-feature inference to the test; however, inference from available information was not among them. In fact, when White and Koehler described the work of Ganzach and Krantz (1990) as demonstrating that missing cues are replaced by mean values, they neglected to consider the latter's Experiment 2, which demonstrated that mean inference does not extend to situations in which predictor cues are intercorrelated. When such is the case, people appear to infer a value for the missing cue based on the values of the other cues with which it is correlated. For instance, the participants in Ganzach and Krantz's Experiment 2 were presented with intelligence and motivation scores for a series of students, and were asked to use the scores to predict the grade-point average (GPA) of each student. In one condition, intelligence and motivation were highly correlated with one another. When a student with an unknown intelligence score was presented, participants did not always infer that the student was of average intelligence; rather, the predicted GPA scores indicated that participants inferred an intelligence score roughly matching that of motivation.

This result may have important implications for future models of categorization. As discussed above, no existing computational models appear to deal with missing data in this way: rather than predicting missing feature values based on learned associations with known cues, they apply some sort of blanket remedy such as simply ignoring the unknown feature. However, as experiments in multivariate prediction, the findings of Ganzach and Krantz (1990) are still one step removed from the issue at hand – the use of missing information in object categorization.

Our primary goal in this study is to examine whether higher-level inferential processes indeed play a role in dealing with missing feature values in categorization, in order to inform future computational modeling efforts in the field. An ideal category structure for testing missing-feature inference would involve a critical cue that, though highly correlated with other cues, nevertheless produces a considerable effect by its presence or absence. With that in mind, we devised two experiments involving missing information and the inverse base-rate effect (IBRE).

The IBRE, first described by Medin and Edelson (1988), is a perplexing categorization effect in which an ambiguous transfer stimulus is classified counter to the principles of normative Bayesian reasoning. In a simplified version of the IBRE, participants learn to diagnose two diseases based on the presence or absence of three symptoms. One disease is common, and is characterized by the presence of headaches and dizziness. The other is rare, and is characterized by headaches and nausea. Headache is thus an (I)mperfectly diagnostic symptom, and is symbolized "I;" dizziness and nausea are (P)erfectly diagnostic of the (C)ommon and (R)are diseases, respectively, and as such are referred to as "PC" and "PR." Thus, the common disease has symptoms I + PC, and the rare disease has symptoms I + PR.

After participants are trained in this category structure, they are presented with a set of novel transfer stimuli. While the ambiguous stimulus I + PC + PR is generally categorized as an instance of the common disease, in line with base rates, the equally ambiguous PC + PR results in the opposite pattern – it is most often categorized as rare. Such a response pattern is puzzling: all other things being equal, the most adaptive strategy would be to classify an ambiguous stimulus as a member of the more common category. This reversal is the inverse base-rate effect, and its exact causes have been the subject of some debate (e.g. Kruschke, 2001; Winman, Wennerholm, & Juslin, 2003; Bohil, Markman, & Maddox, 2005; Winman, Wennerholm, Juslin, & Shanks, 2005). For the purposes of the present study, however, the exact reason behind the IBRE's existence is not important. As long as the effect exists, it can be used to study missing-feature inference.

In both of the experiments reported in this article, we used two pairs of IBRE categories. Thus, there were two rare categories and two common categories in each experiment, along with two PC features, two PR features, and two I features. The different versions are denoted numerically – feature PC1, category R1, and so on. The use of a four-category task ensures that there is some degree of variance on all features during the training period: the training stimuli consist of I1 + PC1, I1 + PR1, I2 + PC2, and I2 + PR2, meaning that each I feature is present in only half of the training examplars. In contrast, in a task using only a single category pair, the imperfectly diagnostic feature I would be present in every single stimulus in the training phase (I + PC and I + PR). The four-category approach is quite common in the IBRE literature (e.g. Kruschke, 1996).

Presenting participants with the stimulus ? + PC + PR in transfer ("?" representing an unknown value for the imperfectly diagnostic feature) would result in a unique opportunity for investigating how people deal with missing information in a categorization task. Mean-value inference theory (White & Koehler, 2004) predicts that people will infer that the unknown feature, I1 or I2, is half-present (as explained above, each imperfectly diagnostic feature would have been present in half of all training stimuli). This inference would result in a response pattern somewhere between those of PC + PR and I + PC + PR – perhaps the mean of the two, although other possibilities will be explored in the General Discussion. Alternatively, if people use known object components to make informed inferences about hidden features, the presence of the PC and PR features should lead to an inference that the I-feature is present, as it was consistently paired with both PC and PR during training. Such a tendency would lead to a response pattern approximating that of I + PC + PR.

The design of Experiment 1 was sensitive to the possibility of higher-level inferences based on abstract rules about the stimulus class. For instance, each training stimulus in the IBRE task contains only two present features. From this, participants might derive a rule that each stimulus may only have two present features, and subsequently apply that rule when attempting to determine the value of the unknown feature in ? + PC + PR. As such, we presented a number of novel transfer stimuli (I, PC, PR, PC + PR, and I + PC + PR) before ? + PC + PR so that participants could draw upon experience with stimuli with varying numbers of present features when making a judgment.

Experiment 1

Experiment 1 consisted of a text-based IBRE-type categorization task (e.g. Kruschke, 2001), with an initial training period followed by a transfer phase in which stimuli with novel feature combinations were presented. In addition, at the end of transfer, participants were asked to categorize stimuli following the abstract form ? + PC + PR, in which the presence or absence of the imperfectly diagnostic feature was unknown.

Method

Participants 75 undergraduates from Simon Fraser University, a large institution in Western Canada, participated in exchange for course credit in introductory Psychology classes.

Materials/apparatus The experiment was conducted using E-Prime stimulus presentation software, running on Apple iMac computers. The stimulus set consisted of text descriptions of fictitious birds, which followed a simplified IBRE category structure. The birds differed on six present/absent features: a feather on the head, a group of spots on the body, a set of sharp claws, a shark-like fin, a cluster of spikes, and a long, rounded tail. Feature assignments and species names were counterbalanced across participants.

Procedure Participants were instructed that they would be learning to identify newly discovered Latin American birds as members of one of four different species (Mexican, Chilean, Bolivian, or Peruvian). The experiment began with a training phase in which the four training stimuli, I1 + PC1, I2 + PC2, I1 + PR1, and I2 + PR2, were displayed in blocks according to a 4:1 base rate (see Table 1). Each trial comprised a single screen of yellow text on a black background. At the top was the sentence "A bird has been

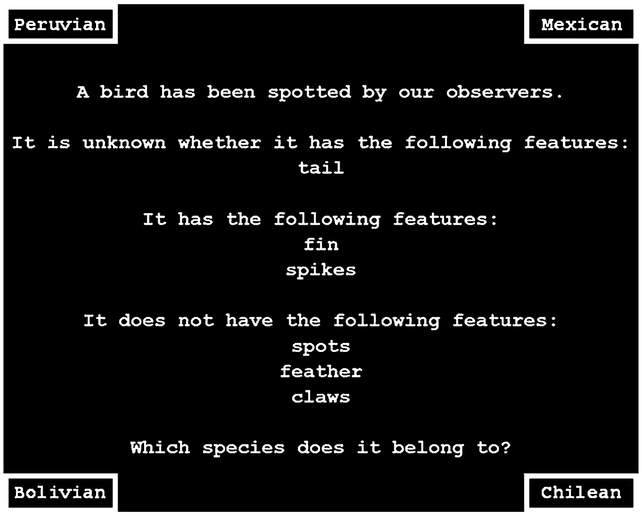
Table 1 Sample category structure for Experiments 1&2

Species	Frequency	Abstract features	Feature names
Chilean	Common	I1+PC1	Fin, tail
Bolivian	Rare	I1+PR1	Fin, spikes
Mexican	Common	I2+PC2	Claws, spots
Peruvian	Rare	I2+PR2	Claws, feather

spotted by one of our observers." Next, missing features were listed (if applicable), then present features, then absent features, each with a descriptive heading (see Fig. 1). The order in which features in each section were listed was randomized across trials.

Participants were able to inspect the descriptions for as long as they wished before using the mouse to click one of the response buttons, located in the corners of the screen. Corrective feedback was provided. The clicked box would turn green following a correct answer; on an incorrect response, it would turn red and the correct answer box would turn green. The training phase continued for 100 trials or until the participant reached a learning criterion of 40 consecutive correct answers, whichever came first.

The transfer phase was split into two stages. The first consisted of standard IBRE transfer stimuli: I, PC + PR, I + PC + PR, PC, and PR. Each stimulus in this phase was displayed twice, for a total of four presentations of each abstract feature combination. In the second stage, participants categorized both versions of the missing-feature transfer stimulus, ? + PC + PR, once each. No feedback was given during transfer; instead, buttons simply turned purple when clicked.



A bird has been spotted by our observers.

It is unknown whether it has the following features:
tail

It has the following features:
fin
spikes

It does not have the following features:
spots
feather
claws

Which species does it belong to?

Fig. 1 Response screen for Experiments 1 and 2, with the stimulus displayed in the middle of the screen and clickable response boxes in the corners. The bird described here is the missing-feature stimulus ? + PC + PR

Results & discussion

Response proportions to transfer stimuli were collected and compared, collapsing across category pairs. Thus, categorizing a bird with features PC1 + PR1 as a member of species C1 was considered to be equivalent to categorizing PC2 + PR2 as species C2. Each transfer stimulus elicited a small proportion of responses inconsistent with the category pair (for instance, categorizing a bird with features PC1 + PR1 as a member of category C2); for the sake of brevity, however, only consistent responses are reported below, and so they do not sum to 100%.

Transfer phase response proportions are shown in Table 2. A considerable IBRE was found, with PC + PR categorized as a member of the appropriate common species 28.0% of the time, and as rare 65.0% of the time. I + PC + PR was categorized more often as common, 60.3%, than rare, 36.0%. Finally, ? + PC + PR was categorized as common 56.7% of the time, and as rare 39.0% of the time. All common-rare differences within these three transfer stimuli were found to be statistically significant (paired-samples *t*-tests, all *p*s < .05). The proportion of common responses to ? + PC + PR was significantly higher than both PC+PR, $t(74) = -6.274$, $p < .001$, and the mean of I + PC + PR and PC + PR (44.2%), $t(74) = 2.804$, $p < .01$. Finally, there was no significant difference in the proportion of common responses to ? + PC + PR and I + PC + PR, $t(74) = 0.731$, $p > .40$, suggesting that participants assumed the unknown feature to be present for the purposes of categorization.

These results are broadly in line with the findings of Experiment 2 of Ganzach and Krantz (1990). The missing feature was reliably correlated with the presence of either PC or PR, and in the presence of both cues people quite reasonably inferred that it, too, was present. The same process appears to be at work in both quantitative prediction and object categorization: people will use available feature values to infer the identity of unknown or missing features, and then use the inferred values to come to a final decision. In Experiment 1, we have extended the findings of Ganzach and Krantz (1990) to the domain of object categorization, demonstrating that current models of category learning are ill-equipped to deal with incomplete information. However, it seems unlikely that predicting features from intercorrelations is the only contributing factor to missing-feature inference; in all probability, there are many situations in which abstract, higher-level reasoning plays a part as well. Experiment 2 seeks to demonstrate an example of such a situation.

Table 2 Distribution of categorization responses to transfer stimuli (Experiment 1)

Stimulus	Consistent common responses (%)	Consistent rare responses (%)	Inconsistent responses (%)
I	79.3	15.0	5.7
I + PC + PR	60.3	36.0	3.7
PC + PR	28.0	65.0	7.0
? + PC + PR	56.7	39.0	4.3
PC	87.3	5.7	7.0
PR	5.0	87.3	7.7

Experiment 2

Experiment 2 is nearly identical in design to Experiment 1, except with one important change: we present the missing-feature stimulus ? + PC + PR immediately after the training phase, and before the standard IBRE transfer stimuli. This results in a comparatively impoverished exemplar space at the time the participant views the missing feature stimulus: all the stimuli experienced up to that point have had exactly two present features. If participants are sensitive to high-level regularities in the stimulus set when inferring values for missing features, they may derive a rule that all stimuli in the set must have exactly two features. Since PC and PR are known to be present in the missing-feature stimulus, a two-feature limit would render the unknown feature necessarily absent and response proportions to the missing-feature stimulus would be identical to those of the transfer stimulus PC + PR. Intercorrelations among features would be just as strong as in Experiment 1, however, so a concordance between ? + PC + PR and PC + PR would indicate that these perceived high-level constraints are even more powerful determinants of participants' inferences than are correlated features.

Method

60 undergraduates at Simon Fraser University participated in exchange for course credit in introductory Psychology classes.

The procedure was identical to that of Experiment 1, with the exception of trial order: the missing-feature stimuli (? + PC + PR) were viewed before the standard IBRE transfer stimuli (PC + PR, I + PC + PR, PC, PR, and I).

Results & discussion

In the standard IBRE transfer phase, responses were largely in line with the results from Experiment 1 (see Table 3). I + PC + PR was categorized as a member of the appropriate common species 62.5% of the time and as a member of the rare species 32.1% of the time. By contrast, PC + PR elicited 32.1% common categorizations, compared to 60.0% rare. Finally, ? + PC + PR was judged to be a member of the appropriate common species 37.5% of the time, and a member of the rare species 52.9% of the time. Differences between common and rare responses within each of these three transfer stimuli were all found to be significant (paired-samples *t*-tests, all *p*s < .05). In addition, I + PC + PR was categorized as common significantly more often than was PC + PR, $t(59) = 5.684$, $p < .001$. ? + PC + PR was categorized as common significantly less than I + PC + PR, $t(59) = 4.924$, $p < .001$, but was statistically no different from PC + PR, $t(59) = 1.217$, $p > .20$. However, ? + PC + PR elicited fewer common responses than the mean of PC + PR and I + PC + PR (47.3%), $t(59) = 2.393$, $p < .05$.

While participants in Experiment 1 appeared to infer that the unknown feature was present, the opposite seems to be true here: people responded to the incomplete stimulus as though the unknown feature were *absent*. The ordering of transfer stimuli, being the only change between Experiments 1 and 2, has radically altered the course of missing-feature inference. This result is predicted by neither mean-substitution theory

Table 3 Distribution of categorization responses to transfer stimuli (Experiment 2)

Stimulus	Consistent common responses (%)	Consistent rare responses (%)	Inconsistent responses (%)
I	71.3	21.3	7.4
I+PC+PR	62.5	32.1	5.4
PC+PR	32.1	60.0	7.9
?+PC+PR	37.5	52.9	9.6
PC	80.8	6.7	12.5
PR	3.3	86.7	10.0

nor the correlated-predictors special case described by Ganzach and Krantz (1990). In fact, an inference of absence is the opposite of what one would expect – in any training stimulus containing either PC or PR, the appropriate imperfectly diagnostic feature was always present. Assuming that the unknown feature is absent is in direct opposition to the feature associations learned during the training phase.

What could cause such a counterintuitive inference? We propose that over the course of training, participants developed a sensitivity to a higher-order regularity of the stimulus class – namely, that each bird had exactly two present features. This was a reasonable assumption for participants to make, as every training stimulus had exactly two present features; at the time of the missing-feature transfer phase, there had been no variance in the number of features a bird could possess. Birds with one or three features did not appear until the standard IBRE transfer phase, which in Experiment 2 came after the missing-feature stimuli. Thus, when presented with a bird known to possess two features with a third of indeterminate value, participants drew upon the two-feature-only rule and assumed the unknown feature to be absent.

General discussion

The results of Experiments 1 and 2 indicate that the process of missing-feature categorization is a good deal more complex than previously suspected. Instead of simply inferring the mean value for an unknown stimulus feature, it appears that people engage in a complex inferential reasoning process to come to a decision about its identity. They take into account correlations with other features, using known values to predict unknown ones and subsequently using the inferred values to generate an appropriate category response. Beyond that, they also draw on experience with the relevant stimulus class to make broad generalizations about what ought to be possible, such as the number of features that can be present. To our knowledge, in spite of the importance of being able to make decisions on the basis of limited information, no current computational model of object categorization accounts for either of these effects. One potential formalization of missing-feature inference from intercorrelated cues comes from the *named error* model of Ganzach and Krantz (1990). In this model, judgments about a numerical outcome (say, predicting GPA from intelligence and motivation scores) are made as a result of a multiple-regression equation, with provided cues as predictors. The mean of each variable in the equation is set to zero for simplicity, and each cue is accorded a particular slope. Inferring the mean for a missing feature, then, is essentially equivalent to setting its value to zero, taking the term out of the equation entirely and moderating the predictions of the model. A special case occurs when the predictors are intercorrelated; in this case, the value of a missing predictor is inferred on the basis of the known cues, and the inferred value is then used as usual to predict an outcome. Ganzach and Krantz did not provide an explicit explanation of exactly how this informed inference process takes place, saying only that an extreme value for a known cue will be reflected in a correspondingly extreme inferred value for a strongly correlated unknown cue.

Rather than presenting intercorrelated predictors as a special case, it seems both more parsimonious and more generalizable to say that the inferred value of an unknown cue is always based on what information is available. Consider a stimulus containing k features in which the value of one feature, X_u, is unknown. We could infer the value of X_u from the available cues using multiple regression:

$$X_u = B_1X_1 + B_2X_2 + \ldots + B_{u-1}X_{u-1} + B_{u+1}X_{u+1} + \ldots + B_kX_{k.}$$

When all other features X_i are uncorrelated with X_u, each slope B_i=0, and the inferred value for the missing feature X_u equals zero – defined as its mean value over all previously stored exemplars. This formulation of missing-feature inference adequately explains empirical results in both correlated- and uncorrelated-cue situations without recourse to special cases, and could be implemented in connectionist models of categorization using a set of weighted, gated connections between input nodes.

While Experiment 1 has demonstrated that intercorrelations among features are useful in missing-feature inference, Experiment 2 showed that other mechanisms are also at work. The response patterns suggest that people decided on the identity of the unknown feature in ? + PC + PR as though making an assumption about the stimulus class constructed during the training phase – namely, that birds have exactly two present features. Any model that hopes to explain the results of both experiments must contain some competitive mechanism to determine the relative importance of these different influences. Whatever that mechanism might be, it is clear that, at least in the present case, regularities across the stimulus class have a powerful effect on responding. The responses to ? + PC + PR were statistically identical to either I + PC + PR (in Exp 1) or PC + PR (in Exp 2), rather than an in-between response as one might expect from a compromise between different influences. Nevertheless, it seems possible that these mechanisms might jointly determine responding, and in other circumstances stimulus class information might have a less overwhelming effect. The role of these kinds of higher level regularities and their underlying cognitive mechanisms are attractive topics for future research.

Alternative explanations may exist for the observed effects. There is some evidence that “missingness” can be informative in itself: Jaccard and Wood (1988) demonstrated that missing information in the description of an alternative makes it somewhat less attractive. It could thus be argued that the apparent inference of absence in Experiment 2 is the result of treating exemplars with missing features as though they were discrepant from all existing exemplars without missing features. If this provision for dealing with missing information were added to the ELMO model, for instance (Juslin, Winman, & Wennerholm, 2001), presenting ? + PC + PR immediately after training would result in a very similar response pattern to PC + PR, since both transfer stimuli differ on the value of the I-feature from the relevant stored exemplars I + PC and I + PR. ADDCOVE (Verguts et al., 2004), or a modified version with the ability to model the IBRE accurately, would react in a similar way, treating ? + PC + PR as a partial mismatch to both training exemplars. This approach, called the *unknown-diagnostic method* by White and Koehler (2004), predicts that regardless of where the mean value of a particular feature lies, a stimulus with that feature missing will be categorized the same way. White and Koehler carried out just such a manipulation in their Experiment 3 by varying the distribution of feature values and found the prediction of the unknown-diagnostic method to be false. Moreover, a model incorporating the unknown-diagnostic method would likely be unable to account for the results of our Experiment 1: ? + PC + PR would be treated as an equal match to PC+PR and I + PC + PR (one discrepant feature, two matching), rather than being categorized identically to the latter.

A potential concern stems from the fact that the previous literature on mean inference as a strategy for dealing with missing data has focused on continuously valued dimensions for which calculating a mean value is sensible – grade point average, for instance. Here, however, the features were binary-valued, making the calculation of a “mean” value somewhat problematic. In the absence of a definitive prediction for the binary-valued case, we reasoned that mean-inference theory would predict a response pattern resembling the midpoint between the absent and present cases. This encompasses two distinct possibilities: first, that the feature is always inferred to be “half-present” (assuming, of course, that a half-present feature leads to a response pattern roughly at the midpoint between absence and presence); and second, that it is inferred to be present on half of the trials and absent on the rest. It is not out of the question that binary-valued predictors constitute a special case; however, the idea that missing-feature inferences are informed by intercorrelations among features even with continuously valued dimensions is further supported by Experiment 2 of Ganzach and Krantz (1990).

Applicability of existing computational models

The results of both experiments appear to run counter to the assumptions of dominant models regarding missing data in categorization. If presented with the stimulus ? + PC + PR, SUSTAIN (Love et al., 2004) would remain agnostic about the value of the missing feature; ALCOVE (Kruschke, 1992) would treat it as a match to both PC + PR and I + PC + PR; and ADDCOVE, a model specifically designed to account for missing data, would treat it as a match to neither (Verguts et al., 2004). In none of these cases is it evident that an inference of presence or absence would result; indeed, it seems most likely that each strategy would result in response proportions falling somewhere between I + PC + PR and PC + PR, with the ordering of transfer stimuli having little or no effect.

Testing the actual performance of the above models is problematic due to differences between the tasks for which some of the models were designed and the IBRE paradigm used in the present experiment. SUSTAIN, for instance, was developed to model tasks such as those used by Shepard, Hovland, and Jenkins (1961) in which there exists a significant amount of within-category variation. In contrast, in the training phase of the IBRE task, the mapping of exemplars to categories is one-to-one: four exemplars, four categories. In this situation of zero within-category variation, SUSTAIN accords all dimensions an equal amount of attention over the course of the training phase (Love et al., 2004), which ultimately cripples the model’s ability to accurately model responses to IBRE transfer stimuli. Likewise, ADDCOVE does nothing to mitigate its predecessor ALCOVE’s inability to accurately model effects such as the IBRE (Verguts et al., 2004; Kruschke, 1992). EXIT, a model which has proven to model the IBRE extremely well (e.g. Kruschke, 2001) is a better candidate. A short description of the model, and of the challenges for its implementation and usage in the context of the present study, follows.

EXIT is a connectionist model with exemplar-specific attentional learning. Activation propagates from the input nodes through to an exemplar-comparison module, which determines the attentional weights used to translate the input pattern into output probabilities. When information about a stimulus is missing, the usual remedy is to simply exclude the unknown features from comparison in the exemplar module. This is the approach assumed to be taken by ALCOVE (Verguts et al., 2004). In EXIT, however, the exemplar module merely determines how the model attends to the input pattern – we are still left with the issue of how

Table 4 Best fit of MEXIT to transfer phase data from Experiment 1

Stimulus	Consistent common responses (%)	Consistent rare responses (%)	Inconsistent responses (%)
I	67.9	17.9	14.2
I + PC + PR	60.9	36.3	2.8
PC + PR	30.4	61.3	8.3
? + PC + PR	45.2	49.4	5.4
PC	82.0	5.1	12.9
PR	2.5	89.8	7.7

Overall RMSD=2.7321.

activation spreads from an incomplete input to the output nodes. Normally, the activation of a given output node k is determined as such:

$$a_k^{out} = \sum_i w_{ki} \alpha_i a_i^{in}$$

where a_i^{in} is the value of the ith feature (0 if absent, 1 if present), α_i is an attentional weight derived from the exemplar module, and w_{ki} is a learned input-to-category association weight (Kruschke, 2001). When a feature is absent, it contributes no activation to any of the output nodes: $w_{ki}\ \alpha_i\ a_i^{in} = 0$. Therefore, if we were to simply exclude an unknown feature from contributing to the activation of any output nodes, this would amount to the same thing as simply assuming it to be absent, an approach at odds both with previous research (Jaccard & Wood, 1988; Ganzarch & Krantz, 1990; White & Koehler, 2004) and with the empirical data above.

In spite of this discrepancy, EXIT shows the most promise of the aforementioned models for modification to handle incomplete information in an IBRE task. As such, we implemented a modified version of EXIT which instantiates mean-inference theory in dealing with missing information (Ganzach & Krantz, 1990; White & Koehler, 2004).

Mean-inference EXIT

Mean-inference EXIT (MEXIT) constitutes a rather minor departure from the standard EXIT model. MEXIT contains a vector of running means for each object feature and updates it at the end of each trial. Since features in EXIT are coded as 0 when absent and 1 when present, the running means can also be thought of as percentages of trials in which each feature is present. When a feature value is missing, it is replaced with the appropriate mean value for the purposes of both exemplar similarity computation and output node activation.

We ran constrained function-minimization simulations fitting MEXIT to the human data from both experiments. Starting parameters included the default EXIT values as well as the parameter values which best fit the standard IBRE transfer stimuli alone for each experiment. The best fitting solutions are shown in Tables 4 and 5 for Experiments 1 and 2, respectively. As can clearly be seen, mean inference does a poor job of fitting the qualitative pattern of the data. In simulations of both experiments, the missing-data stimulus ? + PC + PR elicited a proportion of consistent common responses nearly equal to the midpoint between the proportions of consistent common responses to PC + PR and I + PC + PR.

The results of the above experiments demonstrate two distinct deficiencies in how current models of categorization deal with missing data: they have no mechanism by which an unknown feature's value can be predicted using the values of known features, and they cannot account for higher-level assumptions about unknown feature values. Perhaps the fundamental error which previous models and theories all commit is the assumption that the same inference will always be made for a particular unknown feature. Whether a feature is inferred to be present, absent, matching, nonmatching, or the mean, or ignored entirely, the remedy fails to take into

Table 5 Best fit of MEXIT to transfer phase data from Experiment 2

Stimulus	Consistent common responses (%)	Consistent rare responses (%)	Inconsistent responses(%)
I	74.7	14.9	10.4
I + PC + PR	64.9	33.4	1.7
PC + PR	35.4	58.0	6.6
? + PC + PR	49.5	46.4	4.1
PC	88.5	3.2	8.3
PR	2.1	91.2	6.6

Overall RMSD=3.7831.

account the rest of the stimulus. We believe this approach to be misguided. Instead, the present experiments point to a dynamic, context-sensitive process in which categorizers make the most informed inferences possible given the available information, the stimulus set, and the structure of the task itself.

In real world decision-making, it is rare to have a full complement of relevant information. Indeed, categories are useful precisely because they enable us to make inferences about the unknown. While computational models of category learning have made great strides in the past three decades, the problem of missing data remains largely unaddressed; as long as this remains the case, the applicability of models to real-world categorization situations will be unnecessarily limited. It is our hope that future research will take into account this important issue.

Author Note M.J.W. is now at the School of Psychology, University of Kent. This research was funded in part by a grant from the Natural Sciences and Engineering Research Council of Canada and support from the Canada Foundation for Innovation. We acknowledge the valued contributions of John Alderete, Kevin Douglas, Richard Wright, and the members of the Cognitive Science Lab at Simon Fraser University, without whose help the present study would not have been possible. Correspondence concerning this article should be addressed to Michael J. Wood, School of Psychology, University of Kent, Canterbury, England, CT2 7NZ (e-mail: mw337@kent.ac.uk).

References

Blair, M., & Homa, D. (2005). Integrating novel dimensions to eliminate category exceptions: When more is less. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 31*, 258–271.

Bohil, C. J., Markman, A. B., & Maddox, W. T. (2005). A feature-salience analogue of the inverse base-rate effect. *Korean Journal of Thinking & Problem Solving, 15*, 17–28.

Ganzach, Y., & Krantz, D. H. (1990). The psychology of moderate prediction I: Experience with multiple determination. *Organizational Behavior and Human Decision Processes, 47*, 177–204.

Jacccard, J., & Wood, G. (1988). The effects of incomplete information on the formation of attitudes toward behavioral alternatives. *Journal of Personality and Social Psychology, 54*, 580–591.

Juslin, P., Wennerholm, P., & Winman, A. (2001). High-level reasoning and base-rate use: Do we need cue-competition to explain the inverse base-rate effect? *Journal of Experimental Psychology. Learning, Memory, and Cognition, 27*, 849–871.

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

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 3–26.

Kruschke, J. K. (2001). The inverse base-rate effect is not explained by eliminative inference. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 27*, 1385–1400.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 25*, 1083–1119.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: A network model of category learning. *Psychological Review, 111*, 309–322.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117*, 68–85.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53–79.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75*(13, Whole No. 517).

Verguts, T., Ameel, E., & Storms, G. (2004). Measures of similarity in models of categorization. *Memory & Cognition, 32*, 379–389.

White, C. M., & Koehler, D. J. (2004). Missing information in multiple-cue probability learning. *Memory & Cognition, 32*, 1007–1018.

Winman, A., Wennerholm, P., & Juslin, P. (2003). Can attentional theory explain the inverse base-rate effect? Comment on Kruschke (2001). *Journal of Experimental Psychology. Learning, Memory, and Cognition, 29*, 1390–1395.

Winman, A., Wennerholm, P., Juslin, P., & Shanks, D. R. (2005). Evidence for rule-based processes in the inverse base-rate effect. *The Quarterly Journal of Experimental Psychology, 58A*, 789–815.