Better Learning With More Error: Probabilistic Feedback Increases Sensitivity to Correlated Cues in Categorization

Daniel R. Little
University of Western Australia and Indiana University

Stephan Lewandowsky
University of Western Australia

Despite the fact that categories are often composed of correlated features, the evidence that people detect and use these correlations during intentional category learning has been overwhelmingly negative to date. Nonetheless, on other categorization tasks, such as feature prediction, people show evidence of correlational sensitivity. A conventional explanation holds that category learning tasks promote rule use, which discards the correlated-feature information, whereas other types of category learning tasks promote exemplar storage, which preserves correlated-feature information. Contrary to that common belief, the authors report 2 experiments that demonstrate that using probabilistic feedback in an intentional categorization task leads to sensitivity to correlations among nondiagnostic cues. Deterministic feedback eliminates correlational sensitivity by focusing attention on relevant cues. Computational modeling reveals that exemplar storage coupled with selective attention is necessary to explain this effect.

Keywords: probabilistic categorization, correlated cues, selective attention, exemplars versus rules

The structure of the world is indisputably correlational: Fast and powerful cars, such as Ferraris, are more likely to be fire-engine red than are humble and small cars. Small birds, such as Japanese quail, are more likely to sing than are large birds. Pieces of furniture that contain springs are more likely to be suitable for sitting on than are uncushioned pieces. Notwithstanding the ubiquity of such real-world correlations, much research to date suggests that people’s ability to acquire correlational information is extremely heterogeneous. In categorization, structure is typically instantiated through between-category correlations that define the categories of interest; for example, being a male peafowl is correlated with having an extravagant fan of iridescent tail feathers; and in some songbirds, being male is correlated with the ability to sing: male birds sing to attract female birds. In addition, categories may be characterized by within-category correlations that represent the co-occurrence of multiple features that potentially cross category boundaries and that are, therefore, nondiagnostic for classification (Chin-Parker & Ross, 2002; Murphy & Wisniewski, 1989). For instance, for some small birds, such as robins, both male and female birds sing to demarcate their feeding areas; in this instance, size of the bird and singing share a within-category correlation that falls across the sex category boundary. In this article, we address the question of what types of correlations people learn when they learn about categories. We are particularly concerned with nondiagnostic feature correlations, such as between the power of a car and its color or between the size of a bird and its propensity to sing—neither of which is relevant to determining whether the object is a car or a bird in the first place. Can people learn those correlations? If so, what are the circumstances that facilitate the acquisition of such correlations, and why?

Whether people have access to nondiagnostic feature correlations is important for at least two reasons: First, knowledge of the statistical properties of the environment facilitates prediction of future events. For example, knowing that being small is correlated with singing permits the correct inference that a willie wagtail sings whereas an emu does not, notwithstanding the fact that both animals can be classified as birds without knowledge of either their size or their vocalizations. Second, the use or nonuse of correlational knowledge provides a particularly strong test for models of category learning because some classes of models predict that correlational knowledge should be accessible, whereas other classes of models do not. For instance, a rule supplies knowledge of the category without providing information about the relation between features (e.g., that small birds are more likely to sing than are large birds). By contrast, an exemplar model potentially provides access to this information because all features are stored for each instance: In consequence, information about bird size and song could be retrieved later and compared to produce the correct inference.

To date, a clear picture of correlated-cue knowledge has failed to emerge, owing in part to the use of widely divergent methods.
Relevant research has been conducted in two broad paradigms, namely, category learning tasks, in which the participant intentionally seeks to learn experimenter-defined categories via error-correcting feedback and category usage tasks, in which the participant must use category knowledge to achieve some other goal. The basic finding is that correlated-cue knowledge can be identified in the latter type of task but not the former (Ahn, Marsh, Luhmann, & Lee, 2002; Chin-Parker & Ross, 2002; McNorgan, Kotack, Meehan, & McRae, 2007; Wattenmaker, 1993).

To foreshadow the remainder of this article, we first review these two types of tasks and highlight differences in selective attention as a possible causal factor that determines whether correlated cues are detected. On this assumption, any manipulation that affects selective attention should also affect the detection of correlated cues; we explore this argument with a probabilistic category learning task based on the notion that probabilistic feedback is likely to broaden people’s attention profile. We then report two experiments that demonstrate that probabilistic feedback leads to sensitivity to nondiagnostic correlated cues. Two models that differ in their underlying representations (i.e., exemplar vs. rule) are then fit to the data. The modeling reveals that correlational sensitivity is the result of a broadening of attention across multiple stimulus features that occurs in response to probabilistic feedback. We find that an exemplar representation is required to capture correlational sensitivity, when present.

**Category Usage Versus Category Learning**

Category usage tasks require the application of category knowledge without explicit instructions to learn the categories’ defining attributes. For instance, people may be asked to predict feature values, or people may be asked to provide typicality ratings of exemplars. Generally, all of these tasks reveal sensitivity to correlated cues. For example, when participants rate the typicality of feature pairs for a particular category, typicality tends to be judged higher if the features are correlated than if they are not (Malt & Smith, 1984). People also become more sensitive to correlations when the task requires processing of all stimulus attributes, such as during feature prediction (Anderson & Fincham, 1996; Chin-Parker & Ross, 2002). Chin-Parker and Ross (2002) trained participants either in a category learning task (people responded with the missing category label) or in a feature inference task (responded with a missing stimulus feature). The tasks differed only with respect to the type of missing information that had to be inferred. Chin-Parker and Ross (2002) found that the feature-inference task engendered greater correlational sensitivity than did the category learning task. In the latter task, people failed to detect correlated cues, and this failure turns out to be a pervasive attribute of category learning. For example, Wattenmaker (1991) instructed participants to learn as much as possible about members of a single category whose four stimulus features involved several nondiagnostic correlations. After training, participants demonstrated little ability to access information about these correlated cues and were aware only of diagnostic features and configurations. Similar results have been reported by Murphy and Wisniewski (1989), who found that the effects of prior knowledge far outweigh experimental training of family resemblance categories with co-occurring features.

Why, then, do people show evidence of correlational knowledge in category usage tasks? Previous explanations either have proposed that usage tasks require an exemplar representation, whereas category learning can involve some form of abstraction (i.e., rules or hypotheses; Wattenmaker, 1991, 1993) or have focused on the fact that the two tasks have different goals (McNorgan et al., 2007). A third explanation was suggested by Thomas (1998), who trained participants to categorize two-dimensional stimuli drawn from bivariate-normal distributions. The distributions for the two categories had different means on the two features but an identical, positive, within-category correlation in one condition and an identical, negative, within-category correlation in another condition. Following standard category learning, knowledge of this correlation was assessed with a feature-prediction task. People exhibited knowledge of the correlation when it was positive but not when it was negative. In the latter case, participants instead appeared to form unidimensional rules that discarded one of the diagnostic features. Thomas (1998) proposed that attention to both features was necessary to demonstrate knowledge of the correlation; when participants selectively attended to only one feature (in which the negative-correlation condition was not detrimental to accuracy), correlational knowledge failed to emerge.

In the current experiments, we explore this tentative idea that selective dimensional attention determines sensitivity to nondiagnostic correlations among cues. In particular, we consider circumstances in which attention to nondiagnostic features can be increased, thus possibly creating an opportunity for the detection of correlational information, even in intentional category learning.

**Selective Attention and Correlational Sensitivity**

In most theories of attention in category learning, it is assumed that people attend to those features that will maximize accuracy (Kruschke, 1992, 2001, 2003; Kruschke & Johansen, 1999; Lamberts, 1995; Nosofsky, 1986). This view is supported by multiple streams of evidence, including the relative ease of learning when fewer cues are diagnostic (Kruschke, 1993; Shepard, Hovland, & Jenkins, 1961), the blocking and highlighting effects (Kruschke, 2003; Kruschke & Blair, 2000), the fact that intradimensional validity shifts (i.e., when the response validities of a cue change but its identity does not) are easier to learn than are extradimensional shifts (i.e., when a different cue becomes valid; Kruschke, 1996), and the eye-gaze tracking experiments that show that eye-gaze direction is correlated with feature validity (Kruschke, Kappenman, & Hetrick, 2005; Rehder & Hoffman, 2005a, 2005b). In addition, one theory of attention-shifting proposes that when participants make a classification error, their attention is briefly and fleetingly shifted from the learned diagnostic features to other features of the current stimulus (Kruschke, 2001; Kruschke & Johansen, 1999). By implication, attention shifting may be particularly prevalent in categorization tasks with probabilistic feedback, in which an item’s category membership cannot be predicted from trial to trial with absolute certainty; only long-term trends can be learned (e.g., “this item belongs to category A 70% of the time”). In consequence, perfect performance cannot be achieved, and participants are forced to accept some level of error throughout the task.

The persistence of error, and the attention-shifting it entails, may explain why participants typically underuse valid cues in
probabilistic categorization and overuse irrelevant cues (see, e.g., Castellan Jr., 1973; Edgell, 1980; Edgell et al., 1996; Kruschke & Johansen, 1999). By implication, if attention is required to detect feature correlations, the use of probabilistic feedback, which is known to increase the prevalence of rapid attention shifts, might in turn enhance correlational sensitivity. \(^1\) Categorization tasks are often characterized by large individual differences in strategies (see e.g., Erickson & Kruschke, 1998; Little & Lewandowsky, 2009; Rouder & Ratchliff, 2004; Thomas, 1998; Yang & Lewandowsky, 2003, 2004). This is particularly relevant in the present case because prior research examining correlational sensitivity has relied on complex stimuli with multiple relevant and nonrelevant correlated cues (see e.g., Chin-Parker & Ross, 2002; Wattenmaker, 1991, 1993). It follows that correlational sensitivity may be a byproduct of other types of responding—for example, a nondiagnostic correlation may be a constituent part of an overarching rule involving all available features. Hence, the following experiments involve a simpler design to permit accurate identification of the cues that drive people’s decisions. Our analyses focus first on identifying individual differences and then on how these differences mediate correlational sensitivity.

**Behavioral Experiments**

**Experiment 1**

The principal aim in Experiment 1 was to examine whether people detect a nonrelevance correlation between two cues when categories were probabilistically reinforced. In the experiment, we compared performance with deterministic and probabilistic feedback on a categorization problem involving four binary features. Three of the features (called X, Y, and Z) were instantiated by circles that were either open or filled (hereafter called shading). The fourth feature (called C) was instantiated by the color common to all circles (red or green). Table 1 summarizes the category structure, and sample stimuli are shown in Figure 1.

The table identifies two diagnostic exclusive-or (XOR) compound cues (i.e., pairs of features; Y & Z and Y & C) that were equally valid but that differed in their intrapair relational similarity. Relational similarity refers to the fact that O!O is considered more similar to the target XM than to XXA, despite the fact that XXA—but not O!O—shares features with the target (Goldstone, Medin, & Gentner, 1991). There is evidence that people favor cues that are relationally similar in categorization tasks (e.g., Little & Lewandowsky, 2009); it follows that in this experiment people may also rely on the relationally similar compound (i.e., two circles varying in shading; Y & Z) in preference to the compound involving features of different types (i.e., color and shading; Y & C). The category space additionally included a nondiagnostic correlation, involving the shading of one of the circles (Z) and involving color (C). (See Table 1). Our goal was to determine whether participants have knowledge about the correlation between Z and C that is independent of any knowledge learned about the diagnostic YZ or YC components. The study thus included a posttraining feature-completion task in which participants predicted the missing color feature, given a category label and a partial stimulus. If participants have knowledge of the nondiagnostic zc correlation, then they should consistently respond with the color that was present throughout training. (Hereafter, we introduce the mnemonic of referring to a valid component in upper case, e.g., YZ and YC, and a nonvalid correlated cue of interest in italics and lower case, e.g., zc). Conversely, if participants are using the diagnostic YZ (or YC) compound without knowledge of zc then the completion responses (when averaged across all items) should match the trained colors only 50% of the time: Without knowledge of the zc correlation, people can only guess the color in which a stimulus was presented, thus leading to chance performance.

**Method**

**Participants.** Forty-seven Indiana University students received partial course credit or $10 AUD (approximately U.S.$7) remuneration for participation. Participants were randomly assigned to the deterministic condition (n = 24) and the probabilistic condition (n = 23).

**Stimuli and apparatus.** The assignment of physical stimulus features to the abstract X, Y, and Z dimensions was counterbalanced across participants. The values of each feature assignment were randomized, so that for example, a filled circle might represent a value of 0 and an open circle might represent 1 or vice versa. The color of the stimulus was always mapped to the C dimension, although the value of the C was randomized for each participant (e.g., red might represent a value of 0 and green might represent 1 or vice versa). Depending on condition, the feedback during category learning was either deterministic or probabilistic (see Table 1).

In both conditions, the base-rate validity, or overall mean probability that a stimulus belonged to category A, P(A), was .5. Category B probabilities were 1 − P(A). The component validities are derived from a linear model in which the stimulus components are the independent variables, the category feedback is the dependent variable, and the validity scores are the coefficients in the linear model. The validity scores are equivalent to the slope of the best fitting regression solution with base rate as the intercept (Kruschke & Johansen, 1999). Equivalently, validity can also be expressed as a deviation from the base validity. For example, in both conditions, averaging the probability of category A when dimension X had a value of 0 yields .5; this average deviated from the base rate by 0, indicating that dimension X was not a valid predictor (hence, the validity of X was 0). The only stimulus components with any validity were the YZ and YC compounds. To illustrate, if features Y and Z are recoded as −1 and +1 and multiplied, the average probability of category A when YZ = 1 would be 1.0 in the deterministic condition and .75 in the probabilistic condition, yielding deviations from base rate of .50 and .25, respectively. Thus, the YZ compound (and likewise the YC compound) had a validity of .50 in the deterministic condition and .25

\(^1\) Attention-shifting in probabilistic environments is often assumed to attenuate over the course of learning as participants become accustomed to an unavoidable level of error (Kruschke & Johansen, 1999). At first glance, this attenuation might appear to mitigate against the detection of nondiagnostic correlated features. However, that attenuation has been shown to be far from complete, and participants can learn relevance shifts in a probabilistic task even after 300 trials (Craig, Lewandowsky, & Little, 2008). Hence, we do not consider attenuation to be problematic in the current experiments.
in the probabilistic condition. Note that the probabilistic feedback reduces the validity of the valid cues. Because only stimuli that preserved the zc correlation were shown during training, the categorization problem was isomorphic to a Shepard et al. (1961) Type II category space in both conditions.

Participants were individually trained on an IBM-compatible personal computer running a MATLAB program developed with the Psychophysics toolbox (Brainard, 1997; Pelli, 1991). Each circle was displayed with a radius of 12 cm, and the entire stimulus subtended a visual angle of approximately 30°. Stimuli were displayed in red or green against a white background.

**Design and procedure.** Following precedent (Kruschke & Johansen, 1999), detailed instructions outlining the probabilistic nature of the task and the desirable level of accuracy (around 65%–70%) were provided in the probabilistic condition at the outset. The deterministic condition was not given a target accuracy.

In both conditions, training consisted of 10 blocks of trials, each involving 5 presentations of the 8 training items in a different random order for a total of 400 trials. On each training trial, participants had to categorize the stimulus into one of two categories, \( F \) or \( J \), by pressing the corresponding keys on the keyboard. (Each stimulus was ostensibly presented as a strand of alien DNA to be grouped into two different species of alien). After a response, feedback (“CORRECT” or “WRONG”) was presented underneath the stimulus. Feedback was generated randomly from the probabilities shown in Table 1 and remained visible for at least 1 s, after which, participants could press the spacebar to advance to the next trial. If participants did not press the spacebar, feedback remained visible for 15 s. In addition to feedback for each trial, at the end of each block, the computer also displayed the percentage correct for the previous block.

The final categorization transfer tests contained 10 presentations of the 16 test items in random order, with a 500 ms blank interval between each response and the next item. Transfer trials were identical to training trials, except that feedback was withheld, and

**Table 1**

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>C</th>
<th>P(A) feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Deterministic</td>
<td>Probabilistic</td>
<td>Deterministic</td>
<td>Probabilistic</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Note. The physical instantiation of the three circle features was randomly allocated to the abstract X, Y, and Z dimensions for each participant. The prefix N along with a dash (“–”) in the P(A) feedback column indicates that an item is a new item during the transfer tests. X, Y, Z, and C refer to the four binary features used in the categorization problem. P(A) refers to the overall mean probability that a stimulus belonged to category A.

**Figure 1.** Example of the stimuli used in Experiments 1 and 2. Note that the C dimension was instantiated as the color of the stimulus indicated by the words green or red. X, Y, Z, and C refer to the four binary features used in the categorization problem.
all of the items shown in Table 1 were presented. Instructions for the feature-completion and transfer tests were presented after the learning phase, ahead of each test.

**Results**

Training performance. Participants in the probabilistic condition appeared to be probability matching (PM) their responses to the feedback probabilities. This type of responding is typical in tasks with probabilistic feedback (Friedman & Massaro, 1998; Myers, 1976; Shanks, Tunney, & McCarthy, 2002; Vulkan, 2000) and can be contrasted with maximizing (i.e., consistently responding with the category that has the higher objective probability that achieves the highest possible performance level). To place performance of the deterministic and probabilistic conditions on an equivalent scale, training performance was summarized relative to the feedback probabilities using relevant precedent (Friedman & Massaro, 1998):

\[
PM_i = \frac{1}{N} \sum_{j=1}^{N} [P(A|j) - R(A|j)]S_i + AD_i, \tag{1}
\]

where \(P(A|j)\) is the probability of receiving feedback for category A, given item j (see Table 1), \(R(A|j)\) is participant i’s proportion of A responses for item j, and \(S_i\) is the signed indicator of item j (i.e., \(S_i = +1\) if \(P(A|j) > .5\) and \(S_i = -1\) if \(P(A|j) < .5\) (see e.g., Friedman & Massaro, 1998). These PM scores were averaged across all items for each participant. The scores were then adjusted, AD_i, by adding +1 to the deterministic condition and +.75 to the probabilistic condition; this adjustment means that for both conditions, a PM score of .75 indicates that the response proportions were identical to the objective probabilities of the probabilistic condition. A PM score greater than .75 indicates overshooting of the training probabilities, with PM = 1.00 indicating perfect maximizing. For both conditions, a score of less than .75 indicates undershooting of the training probabilities (with PM = .50 indicating chance performance, and PM < .50 indicating a reversal of the training probabilities).

To ensure that only motivated and able participants were included in the analysis, we removed participants whose average PM scores in the last two training blocks were less than the chance cutoff of .60 (\(\alpha = .01\), using a binomial distribution with \(N = 160\) and \(P = .5\)). No participants were removed in the deterministic condition, but 7 were removed in the probabilistic condition, leaving 24 participants in the deterministic condition and 16 in the probabilistic condition.

The PM scores of both conditions are shown in Figure 2 (see Panel A). Responding in the deterministic condition quickly reached accurate levels. By the end of training, the probabilistic condition had also reached the objective probabilities.

The clear upward trend indicates that both conditions demonstrated learning across training blocks. In confirmation, paired-sample t tests revealed significantly higher PM scores in the final block than in the first block for the deterministic condition, \(t(23) = 13.59, p < .01\), Cohen’s \(d = 2.83\), and the probabilistic condition, \(t(15) = 5.77, p < .01\), Cohen’s \(d = 1.49\).

Transfer performance. Transfer performance was statistically examined by computing the utilization of each stimulus component for each participant. Utilization is computed in the same way as validity (i.e., deviation from base-rate validity), with the exception that response proportions are used in the utilization calculation, whereas the feedback probabilities are used in the validity calculation.

There were clear individual differences in how people learned the task, with some participants preferring the relevant YZ compound and others preferring the YC compound. Following relevant precedent (e.g., Little & Lewandowsky, 2009; Yang & Lewandowsky, 2004), we conducted a k-means cluster analysis (with \(k = 2\); i.e., equal to the number of relevant compounds) on the utilization scores for the two relevant dimensions for all participants of both conditions, simultaneously. The starting points of the cluster analysis were selected randomly; the final cluster centroids returned by the k means showing the utilization of each of the stimulus components are shown in Figure 3. In the deterministic condition, the two clusters were clearly differentiated by the utilization of the two relevant compounds, YZ or YC. The majority of participants (\(n = 18\)) based their performance on YZ, with utilization significantly greater than zero, \(t(17) = 28.15, p < .01\), Cohen’s \(d = 6.83\). A small number (\(n = 6\)) of participants utilized YC instead, \(t(5) = 3.54, p < .05\), Cohen’s \(d = 1.59\). No other components were utilized in either cluster.

In the probabilistic condition, one group utilized the YZ compound (\(n = 7\)), \(t(8) = 3.02, p < .05\), Cohen’s \(d = 1.07\). The remaining probabilistic participants (\(n = 9\)) utilized the YC component, \(t(6) = 2.62, p < .05\), Cohen’s \(d = 1.07\); however, this cluster also utilized several other stimulus components, including Z, \(t(6) = 11.72, p < .01\), Cohen’s \(d = 4.78\); C, \(t(6) = 2.78, p < .05\), Cohen’s \(d = 1.20\), and YZ, \(t(6) = 10.70, p < .01\), Cohen’s \(d = 4.37\). This suggests that at least in the YC cluster, probabilistic participants spread their attention across multiple stimulus com-

\(^{2}\) Of course, in the deterministic condition, probability matching to the objective deterministic probabilities is identical to maximizing. The only other option is undershooting, which would represent a failure to learn the task.
ponents; however, only the feature-completion data provide clear evidence of correlational sensitivity.

Feature completion. Figure 4 shows each cluster’s performance in the feature-completion task. Results are displayed as the proportion of responses in which C was completed with a value of 1. If participants display knowledge of the \( zc \) correlation, then C should be completed with a value of 0 for Items 1–4 (hence, \( P(C = 1) < .5 \)) and a value of 1 for Items 5–8 (\( P(C = 1) > .5 \)), regardless of the category label that is displayed with the item (see Table 1). Hence, for the following analysis, we averaged responses across the two category labels with which each item was shown. A 2 (condition) × 2 (cluster) × 8 (item) between–within analysis of variance (ANOVA) revealed a main effect of item, \( F(7, 252) = 11.33, p < .01, \text{MSE} = .07, \eta_p^2 = .24 \), indicating an overall trend in responding commensurate with knowledge of \( zc \). In addition, the overarching three-way interaction was significant, \( F(7, 252) = 2.64, p < .05, \text{MSE} = .07, \eta_p^2 = .07 \). No other significant effects emerged, with the largest remaining \( F \) ratio being the main effect of cluster, \( F(1, 36) = 2.04, p < .05 \).

The crucial three-way interaction was followed by two separate analyses for each condition. For the deterministic condition, a 2 (cluster) × 8 (item) between–within ANOVA revealed a main effect of item, \( F(7, 154) = 6.82, p < .01, \text{MSE} = .07, \eta_p^2 = .24 \), and an interaction between cluster and item, \( F(7, 154) = 2.79, p < .01, \text{MSE} = .07, \eta_p^2 = .11 \). Figure 4 shows that this interaction arose because only the YC cluster demonstrated knowledge of the \( zc \) correlation (Figure 4, Panel B). This interaction was not significant in the corresponding analysis for the probabilistic condition, \( F(7, 98) = 0.62, p > .05 \), confirming that the main effect of item in this condition, \( F(7, 98) = \)
Discussion

The principal goal in Experiment 1 was to examine whether probabilistic reinforcement leads to knowledge of a nonrelevant feature correlation. This goal was met. In the probabilistic condition, people revealed knowledge of the nonrelevant correlation on the feature-completion task, irrespective of which compound cue they relied on to do the task. Returning to the example from the outset, this outcome is analogous to classifying a bird by gender on the basis of whether it sings and also learning that irrespective of a bird’s gender, small birds tend to sing more than do large birds. Turning to the deterministic condition, most participants (18 out of 24) used the compound involving features of the same type (i.e., two circles; YZ) rather than using a formally identical compound involving features of two different types (i.e., one circle and color; YC). Thus, when given the choice, people used relational similarity to choose between two valid components. In addition, the majority of participants in the deterministic condition (i.e., those who used YZ) did not demonstrate any knowledge of the nonrelevant \( zc \) correlation in the feature-prediction task. This result was very clear and is consistent with previous studies (Medin, Altom, Edelson, & Freko, 1982; Wattenmaker, 1991, 1993), thus reinforcing the fact that correlational sensitivity is generally absent in category learning.

However, contrary to our initial predictions, a minority of participants in the deterministic condition (6 out of 24) exhibited evidence of correlational sensitivity. It is noteworthy that this sensitivity was limited to participants who used the YC component; this sensitivity is perhaps not altogether surprising, given that people who use YC must attend to C during training. Once attention is shifted to C, detecting that feature’s additional involvement in \( zc \) may be facilitated.

In summary, Experiment 1 revealed that with deterministic feedback, correlational sensitivity only arises if attention is directed to stimulus components that contribute to the correlation. With probabilistic feedback, correlational sensitivity arises regardless of how the stimulus components are used in the categorization training task. In Experiment 2, we aimed to increase the generality of this result by using a category space in which the nondiagnostic correlation is not involved in other diagnostic compounds. This should eliminate any occurrence of sensitivity with deterministic feedback.

Experiment 2

In Experiment 2, we used a stimulus space that contained a single, two-dimensional diagnostic compound (XY) and a single, nondiagnostic correlation \( (zc) \). Unlike in Experiment 1, the constituent dimensions (Z and C) of the nondiagnostic correlation \( (zc) \) did not enter into the diagnostic component (i.e., XY). Instead, the \( zc \) correlation unavoidably contributed to a diagnostic overarching XYZC compound. (It is logically impossible to create a nondiagnostic correlation between two cues in a four-dimensional stimulus space without contributing to another diagnostic component; that is, the two-dimensional compound in Experiment 1 or the overarching correlation among all cues in Experiment 2).

By implication, participants could learn the structure in Experiment 2 either by observing the XOR rule involving X and Y or by using XYZC. The latter is rather complex; a verbal rule would note that any item whose color is 0 (see Table 1; \( C = 0 \) may refer to green or red, depending on counterbalancing) with zero or two circles filled and any item of color 1 with one circle filled belongs to category A; otherwise, the item belongs to category B. Use of XYZC subsumes sensitivity to the \( zc \) correlation: Hence, to establish whether people are sensitive to the nondiagnostic correlation per se or whether they exhibit that sensitivity consequent on use of the XYZC compound, in Experiment 2, we again used a feature-completion test in addition to the conventional transfer test.

The transfer and feature-completion tests were identical to those used in Experiment 1. Use rates on the transfer test index which of the two valid cues participants use. The feature-completion test indexes whether participants have knowledge of the \( zc \) correlation. Participants with knowledge of \( zc \) should consistently respond with the color that was present throughout training. Conversely, if participants are using the diagnostic XY component without knowledge of \( zc \), or if participants are using XYZC, then the completion responses should match the trained colors only 50% of the time: Specifically, if people rely on XY without knowledge of \( zc \), they can only guess the color in which a stimulus was presented, thus leading to chance performance. If people rely on XYZC, their performance on the feature-completion task is inconsistent with training because the XYZC rule has contrasting predictions, depending on the category that is presented along with the partial item. For example, if the item 1007 is presented labeled as category B, the XYZC component mandates that the color be 0; see Item 2 in Table 1. However, the same item labeled as category A should lead to the opposite response (see Table 1 and the stated rule, above). It follows that any above chance responding with the trained color on the feature-completion test is uniquely and unambiguously indicative of knowledge of the nondiagnostic \( zc \) correlation, independent of any involvement of a valid component.

Method

Participants and apparatus. Sixty-nine University of Western Australia students received partial course credit or $10 AUD (approximately U.S.$7) remuneration for participation and were randomly assigned to the probabilistic condition \( (n = 38) \) or the deterministic condition \( (n = 31) \). The surface features of the stimuli and apparatus were identical to Experiment 1.

Design and procedure. The stimuli and category structure are summarized in Table 1. The training phase was identical to Experiment 1. Following training, participants completed the feature-completion test, followed by the transfer test. Following these tests, participants completed a typicality rating task; however, because those results are not diagnostic of correlational sensitivity, we elected not to report them, for the sake of brevity.

Results

Training performance. Eight participants in the probabilistic condition who had PM scores less than .6 (the same chance cutoff as in Experiment 1) were removed from analysis, leaving 30 participants in the probabilistic condition and 31 participants in the deterministic condition. Due to a computer error, 2 participants in
the deterministic condition and 4 participants in the probabilistic condition received only one repetition of the transfer items. Hence, the transfer analysis could not be conducted for these participants; the remaining participant numbers were 29 in the deterministic condition and 26 in the probabilistic condition. Figure 2 (see Panel B) shows the PM scores for these participants. Both conditions demonstrated learning across training blocks, again showing higher PM scores in the final block than in the first block for the deterministic condition, \( t(28) = 12.78, p < .01, \) Cohen’s \( d = 2.42, \) and the probabilistic condition, \( t(25) = 7.00, p < .01, \) Cohen’s \( d = 1.40. \) Overall, training performance in Experiment 2 was commensurate with Experiment 1.

Transfer performance. Transfer performance was again examined by computing use of each stimulus component. We again conducted a \( k \)-means cluster analysis \((k = 2)\) on the utilization scores of both conditions simultaneously. In both conditions, we again found that the recovered clusters corresponded to use of the valid components \((XY \) and \( XYZC; \) see Figure 5).

In the deterministic condition, the \( XY \) cluster demonstrated use of the \( XY \) component, \( t(14) = 33.80, p < .01, \) Cohen’s \( d = 9.03. \) No other component had any nonzero utilization. By contrast, the \( XYZC \) cluster used both the relevant \( XYZC \) component, \( t(13) = 6.72, p < .01, \) Cohen’s \( d = 1.86, \) and the relevant \( XY \) compound, \( t(13) = 2.50, p < .05, \) Cohen’s \( d = 0.69. \) The utilization of \( XYZC \) was greater than the utilization of \( XY, \) Cohen’s \( q = 0.73. \) It is important to note that neither of the deterministic clusters used any component that contained dimension \( C \) (aside from \( XYZC); \) hence, the feature-prediction task can diagnose knowledge of \( C, \) independent of use of any of the constituent stimulus components.

The probabilistic condition likewise revealed two clusters that were tied to the relevant stimulus compounds. In the \( XY \) cluster, participants demonstrated significant use of \( XY, \) \( t(18) = 4.92, p < .01, \) Cohen’s \( d = 1.16, \) and of the \( XYZC \) component, \( t(18) = 2.68, p < .05, \) Cohen’s \( d = 0.63. \) For this cluster, the use of \( XY \) was greater than the use of \( XYZC, \) Cohen’s \( q = 0.39. \) For the \( XYZC \) cluster, several components showed significant use scores: \( XYZC, \) \( t(6) = 7.15, p < .01, \) Cohen’s \( d = 2.92; \) \( XC, t(6) = 2.46, p < .05, \) Cohen’s \( d = 1.00, \) and \( Y, \) \( t(6) = 2.88, p < .05, \) Cohen’s \( d = 1.17. \) The effect of \( XYZC \) was larger than both \( Y, \) Cohen’s \( q = 0.79, \) and \( XC, \) Cohen’s \( q = 0.91. \) The utilization scores of both conditions essentially replicated Experiment 1 with a different category space; that is, in both conditions, participants tended to use one or the other of the relevant compounds. In the probabilistic condition, this use was again accompanied by use of other cues, suggesting that attention was more diffuse in this condition.

Feature completion. Figure 6 shows each cluster’s performance in the feature-completion task. The \( 2 \) (condition) \( \times 2 \) (cluster) \( \times 8 \) (item) between–within ANOVA revealed a main effect of item, \( F(7, 357) = 7.53, p < .01, MSE = .07, \) \( \eta^2_p = .13, \) confirming the general trend to respond in accordance with training across both conditions. This result is qualified by the significant Condition \( \times \) Item interaction, \( F(7, 357) = 2.91, p < .01, MSE = .07, \) \( \eta^2_p = .05, \) which confirms that the effect of item was confined to the probabilistic condition. (No other effects of the omnibus analysis were significant, with the largest \( F \) ratio being the main effect of condition, \( F[1, 51] = 2.49, p > .05). \)

Separate follow-up analyses revealed that within the deterministic condition, none of the effects in a \( 2 \) (cluster) \( \times 8 \) (item) between–within ANOVA were significant, with the largest \( F \) ratio being the

---

**Figure 5.** Aggregate transfer utilization rates for each stimulus component (with 95% confidence intervals) for (A) the deterministic condition clusters and (B) the probabilistic condition clusters of Experiment 2. X, Y, Z, and C refer to the four binary features used in the categorization problem.

**Figure 6.** Average proportion of feature-completion items completed commensurate with training (and 95% confidence intervals) for the deterministic condition (A: XY cluster and B: XYZC cluster) and the probabilistic condition (C: XY cluster and D: XYZC cluster) from Experiment 2. If participants have knowledge of \( C, \) then Items 1–4 should have a proportion of responses in which the color dimensions was completed with value 1 (i.e., \( P(C = 1)\)), less than .5, and Items 5–8 should have a \( P(C = 1) \) greater than .5. The solid trend line indicates performance on items averaged across the category label. X, Y, Z, and C refer to the four binary features used in the categorization problem.
main effect of cluster, F(1, 27) = 2.33, p > .05. By contrast, the parallel analysis in the probabilistic condition yielded a main effect of item, F(7, 168) = 8.43, p < .01, MSE = .07, \( \eta_p^2 = .26 \), but no interaction, F(7, 168) = 0.84, p > .05.

To summarize, in this experiment, neither of the clusters in the deterministic condition displayed knowledge of the nonrelevant zc correlation. By contrast, in the probabilistic condition, regardless of which stimulus component was used during training, both clusters exhibited strong sensitivity to zc.

**Discussion**

Confirming our main hypothesis, the feature-completion data revealed that the probabilistic condition—unlike its deterministic counterpart—was sensitive to the nondiagnostic counterpart—was sensitive to the nondiagnostic zc correlation. That sensitivity could not have arisen as a consequence of reliance on either of the valid components alone; in consequence, it presents clear evidence that people have knowledge of a nondiagnostic correlation among features that is in addition to, and not consequent on, knowledge of valid predictors that is demonstrably used to classify stimuli.

It is important to note that the results are not contingent on the cluster analysis. If the feature-completion scores are analyzed without clustering the data, the effect of item arises only in the probabilistic condition, F(7, 175) = 8.01, p < .01, MSE = .07, \( \eta_p^2 = .24 \); the deterministic condition shows no effect of item, F(7, 196) = 1.68, p > .05. Thus, even when clear individual differences are ignored, the data consistently reveal correlational sensitivity in the probabilistic condition but not the deterministic condition. What, then, explains the sensitivity to nondiagnostic correlations observed with probabilistic reinforcement? Previous research has suggested that knowledge of correlated cues is the outcome of memory for individual exemplars (Wattenmaker, 1991, 1993). This interpretation is also attractive in the present case because any participant who memorized the stimuli would presumably be able to recall those exemplars later for the feature-completion task. To explore the idea that exemplar memory underlies correlational knowledge, we now compare how an exemplar model and a class of rule models handle the key results of Experiments 1 and 2.

**Computational Modeling**

**Overview**

We opted to contrast an exemplar model, the generalized context model (GCM; Nosofsky, 1986; Nosofsky & Johansen, 2000), with two rule models based on the set-of-rules model (SRM) introduced by Johansen and Kruschke (2005). The SRM’s mechanisms overlap considerably with those of the GCM, and SRM differs primarily with respect to what is stored in memory. Whereas the GCM stores all previously encountered exemplars, the SRM stores only the relevant feature-to-category associations. The use of two closely related but conceptually quite distinct models has two advantages: First, it allows exploration of our main theoretical questions—is selective attention important in determining correlational sensitivity and can correlational sensitivity be modeled without reliance on the same correlation for classification of exemplars—without commitment to a specific underlying representation. Second, comparing rule-based variants with exemplar-based variants addresses the hypothesis (Wattenmaker, 1993, 1991) that exemplars are central to correlational sensitivity.

To foreshadow our conclusions, the deterministic condition turns out to be better characterized by a rule-based model, whereas

### Table 2

**Stored Feature-to-Category Associations for the XOR-Based Rule Models**

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Rule YZ</th>
<th>Feature</th>
<th>Rule YC</th>
<th>Feature</th>
<th>Rule XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

**Note.** Missing features are indicated by a dash. X, Y, Z, and C refer to the four binary features used in the categorization problem. The dashes indicate that the items were transfer items; hence, these items were not presented during training and have no associated P(A) feedback value. XOR = exclusive-or.

### Table 3

**Stored Feature-to-Category Associations for the Rule-XYZC Model**

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>X Y Z C</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>X Y Z C</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>X Y Z C</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>X Y Z C</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>X Y Z C</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>X Y Z C</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>X Y Z C</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>X Y Z C</td>
</tr>
</tbody>
</table>

**Note.** X, Y, Z, and C refer to the four binary features used in the categorization problem.
the probabilistic condition turns out to be better captured by an exemplar model. In addition to exemplar representations, the probabilistic condition requires a much broader attention profile (i.e., attention to more than just the relevant cues) than does the deterministic condition. Furthermore, representation of nondiagnostic correlated cues turns out to be necessary in the probabilistic condition but not in the deterministic condition.

### GCM

The GCM is an extension of the Medin and Schaffer (1978) context model and captures the basic tenets of exemplar theory (i.e., storage of all previously encountered exemplars, similarity comparison based on psychological distance, selective attention, and relative choice rule). The GCM has been very successful at capturing benchmark findings in category learning; see Nosofsky and Johnson (2000) for a recent review. In the GCM, a stimulus activates all previously encountered stimuli stored in memory according to

\[ s_{ij} \propto \exp(-cd_{ij}), \tag{2} \]

where the similarity, \( s_{ij} \), between items \( i \) and \( j \) is an exponential function of their distance, \( d_{ij} \), in psychological space (Nosofsky, 1986). The steepness of the exponential function is determined by the specificity parameter, \( c \). Selective attention is implemented by differentially weighting the various stimulus dimensions in the distance equation:

<table>
<thead>
<tr>
<th>Group</th>
<th>GCM</th>
<th>Rule YZ</th>
<th>Rule YC</th>
<th>Rule XY</th>
<th>Rule XYC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BIC</td>
<td>RMSD</td>
<td>BIC</td>
<td>RMSD</td>
<td>BIC</td>
</tr>
<tr>
<td>Experiment 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deterministic YZ</td>
<td>5,634.20</td>
<td>.12</td>
<td>4,918.20</td>
<td>.08</td>
<td>7,982.20</td>
</tr>
<tr>
<td>Deterministic YC</td>
<td>1,565.10</td>
<td>.12</td>
<td>2,531.90</td>
<td>.29</td>
<td>2,063.00</td>
</tr>
<tr>
<td>Probabilistic YZ</td>
<td>3,563.70</td>
<td>.14</td>
<td>3,586.00</td>
<td>.15</td>
<td>3,841.00</td>
</tr>
<tr>
<td>Probabilistic YC</td>
<td>2,565.10</td>
<td>.11</td>
<td>2,674.50</td>
<td>.16</td>
<td>3,018.70</td>
</tr>
<tr>
<td>Experiment 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deterministic XY</td>
<td>3,738.40</td>
<td>.09</td>
<td>3,678.00</td>
<td>.09</td>
<td>6,329.60</td>
</tr>
<tr>
<td>Deterministic XYC</td>
<td>5,491.30</td>
<td>.19</td>
<td>5,507.30</td>
<td>.19</td>
<td>4,525.00</td>
</tr>
<tr>
<td>Probabilistic XY</td>
<td>3,031.50</td>
<td>.14</td>
<td>3,572.60</td>
<td>.22</td>
<td>3,400.20</td>
</tr>
<tr>
<td>Probabilistic XYC</td>
<td>6,477.80</td>
<td>.09</td>
<td>6,664.60</td>
<td>.13</td>
<td>6,981.40</td>
</tr>
</tbody>
</table>

Note. The best-fitting model for each group is identified in bold. X, Y, Z, and C refer to the four binary features used in the categorization problem. GCM is generalized context model. BIC is Bayesian information criterion. RMSD is root-mean-squared deviation.

Table 4

### Aggregated Model Fits to the Deterministic and Probabilistic YZ and YC Groups From Experiment 1 and the Deterministic and Probabilistic XY and XYZC Groups From Experiment 2

<table>
<thead>
<tr>
<th>Group</th>
<th>GCM</th>
<th>Rule YZ</th>
<th>Rule YC</th>
<th>Rule XY</th>
<th>Rule XYC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BIC</td>
<td>RMSD</td>
<td>BIC</td>
<td>RMSD</td>
<td>BIC</td>
</tr>
<tr>
<td>Experiment 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deterministic YZ</td>
<td>5,634.20</td>
<td>.12</td>
<td>4,918.20</td>
<td>.08</td>
<td>7,982.20</td>
</tr>
<tr>
<td>Deterministic YC</td>
<td>1,565.10</td>
<td>.12</td>
<td>2,531.90</td>
<td>.29</td>
<td>2,063.00</td>
</tr>
<tr>
<td>Probabilistic YZ</td>
<td>3,563.70</td>
<td>.14</td>
<td>3,586.00</td>
<td>.15</td>
<td>3,841.00</td>
</tr>
<tr>
<td>Probabilistic YC</td>
<td>2,565.10</td>
<td>.11</td>
<td>2,674.50</td>
<td>.16</td>
<td>3,018.70</td>
</tr>
<tr>
<td>Experiment 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deterministic XY</td>
<td>3,738.40</td>
<td>.09</td>
<td>3,678.00</td>
<td>.09</td>
<td>6,329.60</td>
</tr>
<tr>
<td>Deterministic XYZC</td>
<td>5,491.30</td>
<td>.19</td>
<td>5,507.30</td>
<td>.19</td>
<td>4,525.00</td>
</tr>
<tr>
<td>Probabilistic XY</td>
<td>3,031.50</td>
<td>.14</td>
<td>3,572.60</td>
<td>.22</td>
<td>3,400.20</td>
</tr>
<tr>
<td>Probabilistic XYZC</td>
<td>6,477.80</td>
<td>.09</td>
<td>6,664.60</td>
<td>.13</td>
<td>6,981.40</td>
</tr>
</tbody>
</table>

Table 5

### Average Model Parameters and Standard Deviations for the YZ and YC Groups From the Deterministic and Probabilistic Conditions of Experiment 1

<table>
<thead>
<tr>
<th>Model</th>
<th>Group</th>
<th>( \gamma )</th>
<th>SD</th>
<th>( c )</th>
<th>SD</th>
<th>( w_X )</th>
<th>SD</th>
<th>( w_Y )</th>
<th>SD</th>
<th>( w_Z )</th>
<th>SD</th>
<th>( w_C )</th>
<th>SD</th>
<th>( w_{Cat} )</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCM</td>
<td>D-YZ</td>
<td>7.65</td>
<td>16.79</td>
<td>42.43</td>
<td>31.62</td>
<td>.13</td>
<td>.25</td>
<td>.28</td>
<td>.37</td>
<td>.22</td>
<td>.32</td>
<td>.01</td>
<td>.01</td>
<td>.58</td>
<td>.40</td>
</tr>
<tr>
<td>GCM</td>
<td>D-YC</td>
<td>11.48</td>
<td>11.62</td>
<td>30.04</td>
<td>29.70</td>
<td>.53</td>
<td>.44</td>
<td>.33</td>
<td>.35</td>
<td>.21</td>
<td>.36</td>
<td>.14</td>
<td>.20</td>
<td>.54</td>
<td>.48</td>
</tr>
<tr>
<td>GCM</td>
<td>P-YZ</td>
<td>2.51</td>
<td>2.97</td>
<td>15.22</td>
<td>14.93</td>
<td>.31</td>
<td>.32</td>
<td>.39</td>
<td>.36</td>
<td>.21</td>
<td>.20</td>
<td>.27</td>
<td>.34</td>
<td>.58</td>
<td>.49</td>
</tr>
<tr>
<td>GCM</td>
<td>P-YC</td>
<td>3.08</td>
<td>3.90</td>
<td>11.25</td>
<td>9.87</td>
<td>.38</td>
<td>.41</td>
<td>.13</td>
<td>.27</td>
<td>.37</td>
<td>.41</td>
<td>.24</td>
<td>.31</td>
<td>.48</td>
<td>.42</td>
</tr>
<tr>
<td>Rule YZ</td>
<td>D-YZ</td>
<td>10.82</td>
<td>5.18</td>
<td>10.33</td>
<td>8.21</td>
<td>.41</td>
<td>.30</td>
<td>.28</td>
<td>.35</td>
<td>.33</td>
<td>.34</td>
<td>.36</td>
<td>.37</td>
<td>.33</td>
<td>.32</td>
</tr>
<tr>
<td>Rule YZ</td>
<td>D-YC</td>
<td>20.26</td>
<td>4.43</td>
<td>13.03</td>
<td>11.33</td>
<td>.44</td>
<td>.40</td>
<td>.12</td>
<td>.18</td>
<td>.46</td>
<td>.36</td>
<td>.42</td>
<td>.26</td>
<td>.66</td>
<td>.26</td>
</tr>
<tr>
<td>Rule YZ</td>
<td>P-YZ</td>
<td>3.50</td>
<td>3.60</td>
<td>26.46</td>
<td>26.17</td>
<td>.42</td>
<td>.36</td>
<td>.51</td>
<td>.25</td>
<td>.30</td>
<td>.26</td>
<td>.48</td>
<td>.42</td>
<td>.31</td>
<td>.42</td>
</tr>
<tr>
<td>Rule YZ</td>
<td>P-YC</td>
<td>2.98</td>
<td>3.55</td>
<td>27.92</td>
<td>19.07</td>
<td>.32</td>
<td>.40</td>
<td>.24</td>
<td>.32</td>
<td>.56</td>
<td>.36</td>
<td>.26</td>
<td>.42</td>
<td>.44</td>
<td>.42</td>
</tr>
<tr>
<td>Rule YC</td>
<td>D-YZ</td>
<td>10.37</td>
<td>8.26</td>
<td>8.47</td>
<td>12.90</td>
<td>.72</td>
<td>.31</td>
<td>.29</td>
<td>.31</td>
<td>.54</td>
<td>.34</td>
<td>.53</td>
<td>.52</td>
<td>.55</td>
<td>.32</td>
</tr>
<tr>
<td>Rule YC</td>
<td>D-YC</td>
<td>10.64</td>
<td>7.88</td>
<td>9.36</td>
<td>6.86</td>
<td>.35</td>
<td>.30</td>
<td>.31</td>
<td>.32</td>
<td>.16</td>
<td>.16</td>
<td>.21</td>
<td>.18</td>
<td>.15</td>
<td>.17</td>
</tr>
<tr>
<td>Rule YC</td>
<td>P-YZ</td>
<td>3.39</td>
<td>3.40</td>
<td>11.72</td>
<td>11.33</td>
<td>.47</td>
<td>.39</td>
<td>.47</td>
<td>.41</td>
<td>.47</td>
<td>.41</td>
<td>.40</td>
<td>.30</td>
<td>.45</td>
<td>.39</td>
</tr>
<tr>
<td>Rule YC</td>
<td>P-YC</td>
<td>2.01</td>
<td>3.36</td>
<td>29.65</td>
<td>15.47</td>
<td>.42</td>
<td>.39</td>
<td>.28</td>
<td>.46</td>
<td>.27</td>
<td>.33</td>
<td>.56</td>
<td>.37</td>
<td>.39</td>
<td>.42</td>
</tr>
</tbody>
</table>

Note. The attention parameters listed in the table are given at their values before normalization. The bold figures indicate the best fitting model identified in Table 4 as having the minimum Bayesian information criterion. Group names have been abbreviated as follows: D-YZ = deterministic YZ group; D-YC = deterministic YC group; P-YZ = probabilistic YZ group; and P-YC = probabilistic-YC group. X, Y, Z, and C refer to the four binary features used in the categorization problem. GCM = generalized context model.
Table 6

Average Model Parameters and Standard Deviations for the YZ and YC Groups From the Deterministic and Probabilistic Conditions of Experiment 2

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Model Group</th>
<th>Y</th>
<th>c</th>
<th>w_X</th>
<th>w_Y</th>
<th>w_Z</th>
<th>w_C</th>
<th>w_Cat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>GCM</td>
<td>D-XY</td>
<td>7.09</td>
<td>6.37</td>
<td>7.01</td>
<td>4.80</td>
<td>.67</td>
<td>.35</td>
<td>.55</td>
</tr>
<tr>
<td></td>
<td>D-XYZC</td>
<td>5.63</td>
<td>7.39</td>
<td>14.42</td>
<td>16.98</td>
<td>.54</td>
<td>.40</td>
<td>.47</td>
</tr>
<tr>
<td>GCM</td>
<td>P-XY</td>
<td>3.05</td>
<td>2.65</td>
<td>33.42</td>
<td>44.00</td>
<td>.34</td>
<td>.35</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td>P-XYZC</td>
<td>2.99</td>
<td>2.60</td>
<td>13.29</td>
<td>18.63</td>
<td>.23</td>
<td>.32</td>
<td>.34</td>
</tr>
<tr>
<td>Rule XY</td>
<td>D-XY</td>
<td>15.39</td>
<td>12.12</td>
<td>15.32</td>
<td>9.07</td>
<td>.23</td>
<td>.33</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td>D-XYZC</td>
<td>11.53</td>
<td>10.78</td>
<td>13.29</td>
<td>10.02</td>
<td>.26</td>
<td>.32</td>
<td>.29</td>
</tr>
<tr>
<td>Rule XY</td>
<td>P-XY</td>
<td>1.96</td>
<td>1.74</td>
<td>42.60</td>
<td>30.89</td>
<td>.42</td>
<td>.41</td>
<td>.30</td>
</tr>
<tr>
<td></td>
<td>P-XYZC</td>
<td>7.79</td>
<td>19.05</td>
<td>29.57</td>
<td>25.01</td>
<td>.30</td>
<td>.37</td>
<td>.28</td>
</tr>
<tr>
<td>Rule XYZC</td>
<td>D-XY</td>
<td>15.05</td>
<td>11.88</td>
<td>9.77</td>
<td>19.40</td>
<td>.54</td>
<td>.34</td>
<td>.42</td>
</tr>
<tr>
<td></td>
<td>D-XYZC</td>
<td>13.80</td>
<td>9.55</td>
<td>24.85</td>
<td>49.65</td>
<td>.26</td>
<td>.32</td>
<td>.28</td>
</tr>
<tr>
<td>Rule XYZC</td>
<td>P-XY</td>
<td>30.63</td>
<td>41.94</td>
<td>36.48</td>
<td>62.63</td>
<td>.26</td>
<td>.32</td>
<td>.36</td>
</tr>
<tr>
<td></td>
<td>P-XYZC</td>
<td>13.43</td>
<td>23.51</td>
<td>14.55</td>
<td>12.93</td>
<td>.42</td>
<td>.34</td>
<td>.41</td>
</tr>
</tbody>
</table>

Note. The attention parameters listed in the table are given at their values before normalization. The bold figures indicate the best fitting model identified in Table 4 as having the minimum Bayesian information criterion. Group names have been abbreviated as follows: D-XY = deterministic XY group; D-XYZC = deterministic XYZC group; P-XY = probabilistic XY group; P-XYZC = probabilistic XYZC group. X, Y, Z, and C refer to the four binary features used in the categorization problem. GCM = generalized context model.

\[
d_{ij} = \left( \sum_k w_k |x_{ik} - x_{jk}| \right)^{\gamma},
\]

where \(x_{ik}\) is the value of dimension \(k\) for test item \(i\), \(x_{jk}\) is the value of dimension \(k\) for the stored exemplar \(j\), \(w_k\) is the attention weight for dimension \(k\), \(r\) indicates the distance metric, and \(P\) determines the form of the generalization gradient (\(P = 1\), exponential; \(P = 2\), Gaussian; Shepard, 1987). For all simulations, \(r\) and \(P\) were set to unity; typically, \(r\) is set equal to 1 (i.e., city-block distance), for modeling distances between separable dimensions (Shepard, 1991). We assume that the dimensions of the stimuli shown in Figure 1 are separable.

To permit application of the GCM to feature predictions in addition to conventional classification, we also instantiated the category label as a feature dimension and gave it a value of 0 for category A and a value of 1 for category B. Similarity was computed across this dimension for the feature-prediction task. Because the category label was the to-be-predicted feature, \(F_p\), in the categorization test, it was excluded from the similarity comparison there. Conversely, the to-be-predicted feature, \(C\), was excluded from the similarity comparison in the feature-prediction test. The attention weights, \(w_k\), were normalized based on whatever features were used in the similarity computation. Similarities are converted to response probabilities by applying an exponentiated version of Luce’s choice rule (Ashby & Maddox, 1993; Luce, 1963):

\[
P(F_p = 0|j) = \frac{\left( \sum_{F_p = 0} s_j \right)^\gamma}{\left( \sum_{F_p = 0} s_j \right)^\gamma + \left( \sum_{F_p = 1} s_j \right)^\gamma},
\]

where the response scaling parameter, \(\gamma\), allows responding to vary between PM when \(\gamma = 1\) and maximizing when \(\gamma >> 1\) (Ashby & Maddox, 1993; Nosofsky & Johansen, 2000).

Rule Models

The SRM is identical to the GCM in many respects, but it is assumed that only feature-to-label associations are stored. Here, we implemented two models for each experiment motivated by the utilization; for Experiment 1 we implement two XOR rule-models (hereafter called the Rule YZ and Rule YC), and for Experiment 2 we implemented one XOR rule-model and an XYZC rule-model (hereafter, the Rule-XY and Rule-XYZC models, respectively). It is important to note that like the GCM, these models estimate dimensional attention weights from the data.

Because the only valid stimulus components in Experiment 1 and Experiment 2 were the XOR and XYZC compounds, the feature-to-category mappings were also composed of higher level compounds. For the XOR-based models, the stored associations are shown in Table 2. These stored associations replace the stored exemplar, \(j\), in Equation 3. When a stimulus is presented for classification, all of the stored associations enter into the similarity calculation (see Equation 2). Only the features that are represented and that are not being predicted are used to compute the distance between the stimulus and the stored associations (see Equation 3). These similarities are then converted to a response probability with Equation 4.

Strictly speaking, the associations displayed in Table 2 apply only to the deterministic condition. To fit the probabilistic condition, we assumed that only the most frequent rule-to-category association is stored, thus rendering Table 2 applicable also for the probabilistic condition.\(^3\)

As shown in Table 2, the Rule-XOR models explicitly store the XOR mapping from the relevant compounds to categories A and B; hence, performance during the classification transfer test is governed solely by the application of the XOR rule on YZ or YC for Experiment 1 and on XY for Experiment 2. Accord-

\(^3\) If it is instead assumed that each item contributes to feature storage along with its corresponding category label, the fits of the models remain unchanged.
ingly, in the Rule-YZ and Rule-XY models, participants are expected to be unable to predict the missing C feature during the feature-completion task, as neither Z nor C are explicitly represented in the model. The Rule-YC model does have access to the C feature; however, it is tied to the category label and the value of Y. Feature predictions from the Rule-YC model are therefore driven by the association among Y, C, and the category label rather than knowledge of the YC correlation.

For the Rule-XYZC model, the stored associations are more elaborate than in either the XOR-based models or the GCM. Because the XYZC rule takes into account all four of the stimulus features, a full set of stored associations (i.e., all possible combinations of the stimulus features including those not shown during training) are represented (see Table 3). Note that the table entries do not represent exemplars, as in the GCM, but represent associations between each unique configuration of the features and the associated response. In consequence, the table also contains entries for the novel transfer stimuli that were never shown during training but for which responses are prescribed by the XYZC rule. Commensurate with the demonstrable use of all four dimensions for categorization and on use of X, Y, Z, and C in Experiment 2, this model predicts that performance relies on the category label rather than knowledge of the XOR-based models and the GCM. Because the XYZC rule takes into account all four of the stimulus features, a full set of stored associations along with a full set of generalizations (i.e., all possible combinations of the stimulus features including those not shown during training) are represented (see Table 3). Note that the table entries do not represent exemplars, as in the GCM, but represent associations between each unique configuration of the features and the associated response. In consequence, the table also contains entries for the novel transfer stimuli that were never shown during training but for which responses are prescribed by the XYZC rule. Commensurate with the demonstrable use of XYZC in Experiment 2, this model predicts that performance relies on the category label rather than knowledge of the XOR rule models.

\[ \ln L = \sum_i d_i \ln (p_i) + (n_i - d_i) \ln (1 - p_i), \]

where \( p_i \) is the model’s predicted probability of category A for item \( i \), \( d_i \) is the observed number of A responses made for item \( i \), and \( n_i \) is the number of times item \( i \) was presented.\(^4\)

Three models were fit simultaneously to the transfer and feature-prediction data from each experiment; GCM, Rule YZ, and Rule YC were fit to Experiment 1 and GCM, Rule-XY and Rule-XYZC models were fit to Experiment 2. To account for the different number of parameters between the three contenders, we computed the Bayesian information criterion (BIC; Myung & Pitt, 2004), which adds a penalty term to the log-likelihood based on the number of free parameters and the size of the sample being fit:

\( \text{BIC} = -2 \ln L + k \ln n \)

Parameter Estimation and Model Fitting

Each of the models had five attention weight parameters (i.e., \( w_X, w_Y, w_Z, w_C, \) and \( w_{C\text{old}} \)); however, because only features that are not being predicted enter into the decision process, two of the attention weight parameters were effectively zero for the XOR rule models. The other models have four attention parameters. The models also have a specificity parameter, \( c \), and a response scaling parameter, \( \gamma \). Parameters were estimated for each participant separately, by fitting the models simultaneously to an individual’s categorization and feature-prediction responses. The best-fitting parameters were determined by maximizing the log of the binomial-likelihood as follows:

\[ \ln L = \sum_i d_i \ln (p_i) + (n_i - d_i) \ln (1 - p_i), \]

where \( p_i \) is the model’s predicted probability of category A for item \( i \), \( d_i \) is the observed number of A responses made for item \( i \), and \( n_i \) is the number of times item \( i \) was presented.\(^4\)

\( \text{BIC} = -2 \ln L + k \ln n \)

The binomial coefficient has been omitted from the log-likelihood expression because the binomial coefficient is constant across models and parameter settings.

Figure 7. Aggregated model predictions to the deterministic transfer data (means and standard error bars are shown) from Experiment 1. A: Generalized context model (GCM) fit to the YZ group; B: GCM fit to the YC group; C: Rule YZ fit to the YZ group; D: Rule YZ fit to the YC group; E: Rule YC fit to the YZ group; and F: Rule YC fit to the YC group. Old items are items presented during the training and transfer tests, and new items are presented only during the transfer tests. The best fitting model for each group is highlighted with a rectangular outline. X, Y, Z, and C refer to the four binary features used in the categorization problem. \( P(A) \) refers to the overall mean probability that a stimulus was assigned to category A.
where \( k \) is the number of free parameters and \( n \) is the number of cells that enter into the computation of the \( \ln L \).

Model Comparison

Inspections of the individual model fits revealed clear differences in how the models apply across participants; these variations in model fit buttress our decision to differentiate between clusters of participants in our data analysis. To explore these differences and follow relevant precedent (see e.g., Juslin, Jones, Olsson, & Winman, 2003; Lee & Webb, 2005; Little & Lewandowsky, 2009; Navarro, Griffiths, Steyvers, & Lee, 2006; Rouder & Ratcliff, 2004; Yang & Lewandowsky, 2004), we aggregated the model fits across participants within the groups identified by the \( k \)-means analysis. The model fits and parameters averaged across participants within each cluster are presented in Tables 4, 5 (for Experiment 1 parameters), and 6 (for Experiment 2 parameters), respectively. Model predictions for the transfer and feature-prediction tasks are presented separately for the deterministic condition (see Figures 7 and 8) and the probabilistic condition (see Figures 9 and 10) of Experiment 1 and for the deterministic condition (see Figures 11 and 12) and probabilistic condition (see Figures 13 and 14) of Experiment 2.

Not unexpectedly, participants in the deterministic conditions (with the exception of the 6 participants in the YC cluster in Experiment 1, who are addressed below) were well fit by the rule models derived from their respective utilized components (the table entries in Table 4 are \(-\ln L\) values; hence, smaller values indicate better fit). The YZ cluster from Experiment 1 was best fit by the Rule-YZ model (see Figures 7 and 8, Panel C); the XY cluster and XYZC cluster from Experiment 2 were best fit by the Rule-XY and Rule-XYZC models, respectively (see Figures 11 and 12; Panel C for the XY cluster and Panel F for the XYZC cluster). On the basis of the transfer data, the XOR-based rule models and the GCM are difficult to tease apart. The GCM can effectively mimic an XOR model by shifting attention to the components in the XOR compound; for instance, the GCM’s attention parameters for the fit to the deterministic YZ cluster are tuned almost exclusively to \( Y \) and \( Z \) (see Table 5). The feature-prediction data, however, permit clear differentiation between the models. The GCM predicts knowledge of the missing \( C \) component; the rule models do not. Consequently, the deterministic YC cluster was best fit by the GCM (see Figures 7 and 8, Panel A). By increasing attention to \( C \) during training (compare attention with \( C \) for the deterministic YZ group in Table 5), participants in this condition are able to predict the missing \( C \) feature during the feature-prediction task. Neither the Rule-YZ model nor the Rule-YC model is able to produce this pattern of performance.

For the probabilistic conditions, in all cases there was a clear trend to respond with the missing feature, as expected on the basis of training; furthermore, only the GCM predicted this type of performance (see Figures 10 and 14, Panels A and B). The superior fit of the GCM suggests that the correlational sensitivity revealed

\[
BIC = -2 \ln L + k \ln (n),
\]

(6)
by the feature-prediction task in the probabilistic conditions was driven by memory for the training items. Turning to the parameter values from the GCM, it is also evident that the probabilistic conditions were characterized by a broader attention profile than were the deterministic conditions (see Tables 5 and 6). Taken together, those two aspects of the GCM’s performance provide a theoretical link in support of our central thesis, namely, that the diffusion of attention across multiple dimensions underlies correlational sensitivity when present.

To illustrate this point, consider the normalized attention weights for the best fitting parameters from the YZ clusters and the XY clusters. The deterministic YZ group focused solely on the Y and Z dimensions during the categorization transfer tests ($w_Y = 0, w_Z = .53, w_C = 0$) and on the category label for the feature-prediction test ($w_X = 0, w_V = .40, w_Z = .34, w_{Cat} = .26$). By contrast, in the probabilistic YZ group, attention was spread across all of the stimulus dimensions in both tasks (categorization transfer test, $w_X = .30, w_Y = .31, w_Z = .19, w_C = .20$; feature-prediction test, $w_X = .25, w_Y = .28, w_Z = .15, w_{Cat} = .32$). Similar results were found for the deterministic XY group; attention was focused solely on the X and Y dimensions in the categorization task and also on the category label in the feature-prediction task (categorization, $w_X = .42, w_Y = .58, w_Z = 0, w_C = 0$; feature prediction, $w_X = .30, w_Y = .42, w_Z = 0, w_{Cat} = .28$). The probabilistic XY condition, however, showed a diffused set of attention weights (categorization, $w_X = .31, w_Y = .21, w_Z = .28, w_C = .20$; feature prediction, $w_X = .23, w_Y = .16, w_Z = .13, w_{Cat} = .48$).

Summary of Modeling Results

The modeling results are readily summarized.

1. Examination of fits to individual participants confirmed the presence of different subgroups who utilized different compound cues. In consequence, we fit the three candidate models for each experiment, separately, to the two principal groups in each condition.

2. Not unexpectedly, most of the deterministic groups were best accommodated by the rule models designed for that specific group. This result confirms that attention in the deterministic conditions was focused solely on the stimulus dimensions that entered into the rule. These groups exhibited no knowledge of the nondiagnostic $zc$ correlation.

---

5 Because the deterministic XYZC cluster involves use of a rule that encompasses all of the stimulus dimensions, comparison of the attention weights for this cluster is not useful for present purposes. Likewise, the deterministic YC cluster also shows a diffuse attention profile; however, there were only a small number of participants ($n = 6$) in this cluster. We therefore consider it to be less representative of the deterministic condition.
3. Surprisingly, a small number of participants in the deterministic condition were well fit by the GCM and exhibited attention to all of the stimulus dimensions. These participants demonstrated knowledge of the nonrelevant \(zc\) correlation.

4. It is crucial to note that regardless of group allocation, all of the probabilistic participants in both experiments demonstrated clear knowledge of \(zc\) on the feature-prediction test.

5. This correlational sensitivity could be captured only by the GCM and not by the rule models suggesting that storage of more than just a rule is required to permit the emergence of correlational knowledge.

6. The pattern of parameter estimates in the GCM supported the contention that a broader attention profile is required to model correlational sensitivity.

**General Discussion**

**Summary of Results**

The present article offers several contributions:

1. Three of the deterministic groups from Experiments 1 and 2 (containing 47 out of 53 participants) replicated the well-established finding that people are insensitive to nondiagnostic correlations among cues in category learning (Chin-Parker & Ross, 2002; Medin et al., 1982; Wattenmaker, 1991, 1993).

2. In Experiment 1, we extended previous work (see e.g., Medin et al., 1982; Wattenmaker, 1993) by showing that participants favored conjunctions of cues that were of the same type (i.e., \(Y\) and \(Z\)), in preference to conjunctions of cues of different types (i.e., \(Y\) and \(C\)). This result highlights the fact that relational similarity between cues contributed to participants’ decisions.

3. Despite the preference for relational similarity, a small number of participants receiving deterministic feedback opted to use a conjunction of cues of different types. These participants (6 out of a total of 53 in the deterministic conditions) also demonstrated knowledge of the nondiagnostic correlation.

4. The data showed that probabilistic feedback increases sensitivity to nondiagnostic correlational information. To our knowledge, this result represents the first demonstration of correlational sensitivity in intentional category learning when prior knowledge was not available to guide learning.
5. The modeling of Experiments 1 and 2 went beyond previous research by demonstrating (a) that a diffusion of attention across multiple cues is central to the acquisition of correlational knowledge and (b) that this correlational knowledge is best captured by an exemplar-based model and may not be readily explainable by a rule-based model. Because participants were not given the instructions to the feature-completion task until immediately prior to its commencement, the representations used on the feature-completion test had to be acquired at training. Hence, we suggest that probabilistic feedback leads to exemplar storage that allows access to correlated cues via selective attention.

Connections to Prior Research

In the present experiments, we used probabilistic feedback as a means of encouraging sensitivity to correlational information. Thomas (1998) also assessed sensitivity to correlational information, using a different type of probabilistic task. Thomas used large, overlapping categories formed by drawing continuous-dimensional stimuli from a bivariate-normal distribution. Large, overlapping categories are probabilistic because perfect performance is not attainable even if the decision boundary between categories is optimally placed (because the category A distribution overlaps the boundary, some category A stimuli will fall on the side of category B). In that experiment, correlational sensitivity was tied to the dimensionality of the boundary: Correlational sensitivity was observed when participants used a multidimensional boundary but not when the chosen boundary was unidimensional. The current results extend Thomas’s finding by showing that it is not the boundary position per se but rather the distribution of attention underlying any boundary placement that gives rise to correlational sensitivity. In addition, like the current experiments, exemplar processes have also evinced to explain performance with large, overlapping categories (McKinley & Nosofsky, 1995).

Another interesting case of correlational sensitivity in intentional category learning is knowledge partitioning. In knowledge partitioning, training typically involves stimuli composed of one or two continuous dimensions and a binary context dimension (often instantiated as the color of the stimulus; see e.g., Kalish, Lewandowsky, & Kruschke, 2004; Lewandowsky, Kalish, & Ngang, 2002; Lewandowsky & Kirsner, 2000; Lewandowsky, Roberts, & Yang, 2006; Little & Lewandowsky, 2009; Yang & Lewandowsky, 2003, 2004). Although context by itself does not predict category membership, it reliably identifies which of a number of partial boundaries involving the continuous dimensions is applicable for a given stimulus; that is, context is correlated with a region of the stimulus space and the rules that apply to that region.
The typical finding in these knowledge-partitioning tasks is that about one third of participants use the correlation between context and regions of the category space to break the problem into simpler components. In the current experiments, knowledge partitioning would have been identified by the use of one rule when the stimuli were shown in one color (or context, in knowledge-partitioning terminology) and by the use of a different rule when stimuli were shown in the other color (e.g., responding “A” either when dimension Y was open and the stimulus was green or when Y was filled and the stimulus was red, in Experiment 1). This type of responding would have been characterized by total sensitivity to the correlation between Y and C. Instead, the current experiments demonstrate partial, but not total, sensitivity to the correlation between Y and C. One difference between the current task and the knowledge-partitioning paradigm is that in the current experiments, the number of levels of all dimensions (i.e., X, Y, Z, and context) was equal. From a simplicity perspective (Feldman, 2003, 2006; Pothos & Chater, 2002), this means that dividing the stimulus space into smaller components on the basis of color would have yielded no advantage to dividing the space on the basis of any of the other dimensions. Hence, the use of continuous cues may turn out to be an important factor in the emergence of knowledge partitioning.

The current experiments also point to a future direction for intentional category learning involving prior knowledge. Previous research has shown that correlational sensitivity emerges in deterministic category learning if prior knowledge provides a pointer to the correlational information (Ahn et al., 2002; Hayes, Taplin, & Munro, 1996; Malt & Smith, 1984; Murphy & Wisniewski, 1989). For instance, people are quicker to learn that a large brain is related to better memory than they are to learn that a large brain is related to having a rounded beak (Barrett, Abdi, Murphy, & Gallagher, 1993). Prior knowledge is assumed to interact with current learning either by providing a repository of exemplars that can be used in conjunction with learning in the current task (Heit, 1993, 1994, 2000) or by proliferating prior knowledge that is consistent with exemplars in memory (Heit, 1993, 1998, 2000). Our results suggest that the only way for prior knowledge to trigger correlational sensitivity is by guiding selective attention to correlated features, perhaps through direct action on the attention weights. Intriguingly, the correlation itself need not necessarily be represented in the prior knowledge because diffusion of attention is sufficient to engender sensitivity during experimental learning.

**Theoretical Implications**

The GCM adequately fit several aspects of the probabilistic transfer data, including, in particular, the feature-prediction results. However, because the GCM lacks an endogenous mechanism to control how attention is distributed during learning (parameters are instead estimated from the data), GCM’s account of probabilistic categorization remains incomplete. We therefore suggest that the
GCM provided a straightforward and effective way to link attention to the observed behavior in the experiments without, however, explaining why the attentional diffusion occurs—it remains a task for future theory development to design an attentional learning mechanism that responds to PM in accord with the present data.

From a computational perspective, on any one trial, the problem confronting the learner is how to predict the appropriate category. To solve this problem, the learner must sample features from the stimulus and use these features to generate category predictions. The computational problem is how to sample features in a meaningful way. Prior research suggests that features are primarily sampled on the basis of their saliency; however, given enough time, priority is given to features with higher validity (Lamberts, 1995, 2002). In the present experiments, little is known about the feature validities at first, so participants sample uniformly and accumulate noisy counts (or associations) between the correct category and the value of the stimulus dimension. Feature validity is determined by simply looking at whether values of a particular feature are associated only with one category outcome. In the deterministic condition, counts would quickly accumulate for the relevant features, and the accumulation of validity for the relevant features would outpace the nonrelevant features. As a consequence, the deterministic condition would typically sample only the relevant features. In the probabilistic condition, counts would accumulate for more than just the relevant features due to the noisy feedback. Consequently, participants in the probabilistic condition would tend to sample from a wider array of features.

It is interesting to note that to access correlational knowledge, such an approach requires that participants update counts for feature co-occurrences rather than only between each feature and the category outcome. As noted by Anderson and Matessa (1992), fully updating all of the feature-to-feature correlational distributions requires considerable computational costs. One way to deal with the added cost of updating feature co-occurrences is to update the co-occurrence only if the features are sampled together. The GCM does this naturally by shifting attention (cf. Anderson & Fincham, 1996).

**Practical Implications**

The computational analysis outlined above makes our results particularly relevant for studies of cross-situational word learning and statistical learning (Smith & Yu, 2008; Yu, 2008; Yu & Smith, 2007). Researchers in these contexts have used simple associative models as an explanation for how people (children in particular) learn mappings (i.e., correlations) between words and their referents (see e.g., Yu, 2008). The world in which children learn is highly probabilistic—words are not always or even typically paired with their referents in a perfect or optimal fashion. For example, parents may refer to an object in the living room variously as a couch or as a sofa (or even as a chaise lounge), and the family’s kitty may variously be called a cat, a feline, or a pussycat. For many decades, this seemingly suboptimal structure of the learning environment has attracted much theoretical concern and
bewilderment at children’s seemingly effortless ability to acquire language. Indeed, this type of probabilistic learning environment has been deemed to be too impoverished to permit learning without the assistance of a specialized and innate device (Chomsky, 1965). Our data, by contrast, show that probabilistic reinforcement can be a positive feature of a learning environment because probabilistic reinforcement permits the extraction of correlations, even if they are not immediately relevant to the task at hand. In a nutshell, our data suggest that people can learn more when the environment is worse.

The finding that people who undergo probabilistic training have additional knowledge that manifests whenever the task problem changes (i.e., from categorization to feature-prediction) is related to prior work on the study of skill transfer in applied settings (see Lewandowsky, Little, & Kalish, 2007, for a review). Many of the transfer errors that occur on-the-job result either from (a) an inability to recognize similarities between two structurally similar problems (usually due to a mismatch on surface features; Dienes & Altmann, 1998; Gick & Holyoak, 1983, 1980) or (b) a misapplication of previously learned skills to a new task that requires a different approach (Hershey & Walsh, 2000). From a processing point of view, both of these errors can be explained by appealing to a focus of attention during the training task. Our results suggest that probabilistic reinforcement during training (i.e., demonstrating what typically works rather than what always works) might result in attention to more than just a few features, thus allowing knowledge about the entire problem structure to be available for new situations. Of course, another solution would be to train people on multiple tasks; however, probabilistic feedback might prove to be more efficient in some circumstances.

**Conclusion**

The current experiments revealed that category learning in a probabilistic environment results in the acquisition of more statistical knowledge about the category space than does learning with deterministic feedback. Put into everyday terms, this finding is equivalent to learning more nondiagnostic feature correlations about a natural category (e.g., that small birds also tend to sing), if on occasion one encounters members of a different category that also preserve that correlation (e.g., small children tend to sing as well). We have proffered an account of this seemingly paradoxical finding based on a diffusion of dimensional attention resulting from the introduction of probabilistic feedback.

**References**


Received August 23, 2008
Revision received December 21, 2008
Accepted February 20, 2009 ©