Journal of Experimental Psychology:
Learning, Memory, and Cognition
2005, Vol. 31, No. 5, 811–829

Copyright 2005 by the American Psychological Association
0278-7393/05/$12.00 DOI: 10.1037/0278-7393.31.5.811

Thirty-Something Categorization Results Explained: Selective Attention, Eyetracking, and Models of Category Learning

Bob Rehder and Aaron B. Hoffman
New York University

An eyetracking study testing D. L. Medin and M. M. Schaffer's (1978) 5–4 category structure was conducted. Over 30 studies have shown that the exemplar-based generalized context model (GCM) usually provides a better quantitative account of 5–4 learning data as compared with the prototype model. However, J. D. Smith and J. P. Minda (2000) argued that the GCM is a psychologically implausible account of 5–4 learning because it implies suboptimal attention weights. To test this claim, the authors recorded undergraduates' eye movements while the students learned the 5–4 category structure. Eye fixations matched the attention weights estimated by the GCM but not those of the prototype model. This result confirms that the GCM is a realistic model of the processes involved in learning the 5–4 structure and that learners do not always optimize attention, as commonly supposed. The conditions under which learners are likely to optimize attention during category learning are discussed.

Keywords: categorization, eyetracking, prototype models, exemplar models

For almost 3 decades, there has been a debate regarding whether human category learning is best described by prototype or exemplar models. According to the prototype view (Hampton, 1979; Nosofsky, 1992; Rosch, 1975; Rosch & Mervis, 1975; Smith & Medin, 1981), classification is based on similarity to a summary representation. According to the exemplar view (Estes, 1994; Kruschke, 1992; Nosofsky, 1986; Medin & Schaffer, 1978), classification is based on similarity to memory traces for previously classified items. Deciding between these competing hypotheses has proven to be surprisingly difficult.

In a classic study designed to distinguish prototype and exemplar theories, Medin and Schaffer (1978) constructed the now infamous 5–4 category structure listed in Table 1. This category structure represents stimuli that vary along four binary-valued dimensions and make up two opposing categories, A and B. It was designed such that the prototype and exemplar theories make different qualitative predictions regarding the classification of certain items. Consider the classification of Items A1 and A2 from Category A in Table 1. According to prototype theory, the prototype of Category A is 1111, because 1 is the most common value on each dimension in Category A, and Category B's prototype is 0000, because 0 is the most common value on each dimension in Category B. Thus, because A1 (1110) shares three features with the A prototype but only one with the B prototype, the prototype model predicts that A1 will be classified as a member of Category A more often than will A2 (1010), which shares two features with both prototypes. In other words, the prototype model predicts an A1 advantage over A2 in classification performance. The exemplar model, in contrast, predicts an A2 advantage, because whereas A2 shares three features with two Category A members (A1 and A3) and two or fewer with any Category B member, A1 shares three features with only one other Category A member (A2) and two Category B members (B1 and B2). In fact, Medin and Schaffer (1978) found that participants were more likely to classify Exemplar A2 than A1 as an A—that is, they found an A2 advantage favoring the exemplar model. Because the exemplar model also provided a better quantitative fit than the prototype model to the participants' binary responses, Medin and Schaffer's study has been cited widely as evidence in favor of the exemplar view over the prototype view. Because of its theoretical importance, the 5–4 problem has been replicated at least 30 times between 1978 and 2000 (see Smith & Minda, 2000, for a full list of references) and a few times since (Johansen & Palmeri, 2002; Minda & Smith, 2002).

Recently, however, Smith and Minda (2000; Minda & Smith, 2002) have challenged the exemplar model's apparent superiority in accounting for the 5–4 problem. First, they noted that the exemplar model's quantitative superiority might have arisen because researchers fitted aggregate group data rather than individuals—that is, the evidence that the 5–4 categories were represented in terms of exemplars may be an artifact of averaging and may not reflect any one individual's concept formation strategy. To substantiate this claim, Minda and Smith (2002) conducted yet another replication of the 5–4 task and fitted the two models to individual participants rather than to aggregate data. They found that the prototype model indeed produced reliably better fits than the exemplar model. Consequently, Minda and Smith concluded that the dominance of the exemplar view had been based, in part, on fitting aggregate data rather than individual participants.

Although they have conceded that models should be fit to individual learners rather than groups, exemplar theorists have been critical of the fitting procedure used by Minda and Smith

Bob Rehder and Aaron B. Hoffman, Department of Psychology, New York University.

We thank Lewis Bott and Gregory L. Murphy for their comments on previous versions of this article.

Correspondence concerning this article should be addressed to Bob Rehder, Department of Psychology, New York University, 6 Washington Place, New York, NY 10003. E-mail: bob.rehder@nyu.edu

Table 1
The 5–4 Category Structure

	Dimension diagnosticity			
Stimulus	High 1	Low	High 2	Medium
Category A items				
A1	1	1	1	0
A2	1	0	1	0
A3	1	0	1	1
A4	1	1	0	1
A5	0	1	1	1
Category B items				
B1	1	1	0	0
B2	0	1	1	0
B3	0	0	0	1
B4	0	0	0	0
Transfer items				
T1	1	0	0	1
T2	1	0	0	0
T3	1	1	1	1
T4	0	0	1	0
T5	0	1	0	1
T6	0	0	1	1
T7	0	1	0	0

(2002). Nosofsky and Zaki (2002) argued that Minda and Smith used a restricted version of the generalized context model (GCM), one that was constrained to probability match, that is, to predict binary choice probabilities that match the relative ratio of the test exemplar's summed similarity to the two categories. However, it is now generally accepted that category learners often respond in a manner that is more deterministic than probability matching (Ashby & Gott, 1988; McKinley & Nosofsky, 1995). Nosofsky and Zaki also noted that because there is an implicit flexibility in the multiplicative prototype model (MPM) that allows it to model deterministic responding, the GCM was placed at a disadvantage by Minda and Smith's attempts to equate the two models with respect to free parameters. Zaki, Nosofsky, Stanton, and Cohen (2003) found that once the GCM was equipped with a parameter (gamma) that allowed it to model more deterministic responding, it provided better fits than the prototype model to Minda and Smith's (2002) 5–4 data.

The second challenge posed by Minda and Smith (2002) is a theoretical argument concerning how category learners allocate attention. An important characteristic of the 5–4 problem is that the four stimulus dimensions are not equally useful for discriminating the two categories. Whereas three of the four dimensions provide substantial evidence of category membership (two independently predict the correct category seven out of nine times, and a third does so six out of nine times), the fourth only predicts the correct category five out of nine times. In fact, the 5–4 structure's optimal attention weights for both the GCM and the MPM (i.e., the weights that most closely yield perfect classification performance on the nine training items) are .35, .35, .30, and .00 for the two highest, the medium, and the least diagnostic dimensions, respectively, for the MPM, and .33, .33, .31, and .02 for the GCM (Lamberts, 1995).[1] That is, regardless of whether the 5–4 categories are represented as prototypes or exemplars, an optimal classifier should virtually ignore the least diagnostic dimension. In fact, when the MPM is fit to 5–4 human data, the attention weights it estimates suggest that most learners indeed place minimal weight on the least diagnostic dimension. However, when the GCM is fit to the same data, its estimated attention weights suggest the opposite: that learners place substantial weight on the least diagnostic dimension. That is, the models make qualitatively different claims regarding whether people learn to optimize attention while acquiring this category structure: The MPM fits suggest they do; the GCM fits suggest they do not. Minda and Smith argued that this suboptimal allocation of attention suggested by the GCM but not the MPM should be interpreted as evidence that the MPM is, in fact, the more accurate psychological model of performance on the 5–4 problem.

Why should the suboptimal allocation of attention be taken as evidence against the GCM? One reason is that the exemplar model has frequently been accompanied by suggestions that category learning includes learning to attend optimally (Lamberts, 1995; Nosofsky, 1984). Another is that attention learning has been explicitly incorporated into the connectionist implementation of the GCM known as ALCOVE (Kruschke, 1992). Indeed, attention learning is at the heart of ALCOVE's ability to account for some of the classic data sets in the literature. For example, Shepard, Hovland, and Jenkins (1961) presented participants with a number of category structures involving three binary dimensions that varied in the number of dimensions needed for perfect performance: one dimension (the Type I structure), two dimensions (the Type II structure), or all three (Types III–VI). The well-known result is that learning was influenced by the number of relevant dimensions (the ordering of difficulty was I < II < [III, IV, V] < VI). In fact, ALCOVE reproduced these findings in part by learning to ignore the two irrelevant dimensions in the Type I structure and the single irrelevant dimension in the Type II structure (Kruschke, 1992). However, if Shepard et al.'s participants optimized attention on the Type I and Type II problems, then why did the 5–4 participants fail to optimize (at least according to the GCM fits)? For Minda and Smith (2002), this inconsistency raised the question of whether those GCM fits were veridical—that is, whether they provided a psychologically realistic model of learners' performance on the 5–4 problem despite its superior quantitative fit.

To determine whether the attention weights estimated by the exemplar model were psychologically realistic, Minda and Smith (2002) attempted to independently assess those weights by calculating correlations between individual category responses and the abstract values on each dimension. In fact, they found small correlations between category responses and the least diagnostic dimension, a result that was consistent with the prototype model's low average attention weight estimates for that dimension and inconsistent with the exemplar model's high weight. However, Zaki et al. (2003) demonstrated that there are problems with evaluating attention by this method. First, because the analysis assumes independent dimensions (consistent with the prototype

[1] These optimal weights were computed under the assumption that *sensitivity* (i.e., the c parameter) was equal to 3. Note that because the optimal weights depend on the value of c, there is no unique set of optimal weights. For example, if c is very high, almost any attention profile will yield perfect performance, whereas the weights become more important at lower c values. Nevertheless, over a large range of psychologically realistic values of c, the optimal weights indicate a qualitatively similar pattern.

model) and thus ignores configural information (contrary to the exemplar model), it is biased to confirm the prototype model's estimated attention weights. Second, Zaki et al. (2003) simulated response probabilities from hypothetical categorizers with evenly distributed attention across the four dimensions. A correlational analysis incorrectly concluded that these hypothetical categorizers ignored the low dimension. In other words, a correlational analysis does not provide the independent assessment of participants' attentional allocation that Minda and Smith (2002) hoped it would.

Nevertheless, the question of how learners allocate attention for the 5–4 problem remains. One reason this question is important is that (as just discussed) it speaks to whether people always optimize attention while learning categories. However, another reason is that it provides a decisive test of the GCM and the MPM. Whereas these models make very similar predictions for learners' classification choices (they can be distinguished only by small differences in the probability of classifying certain category items; e.g., A1 and A2), they imply a large difference in how learners allocate attention. Thus, the question of whether the 5–4 problem is learned through prototype abstraction or exemplar memorization could be more definitively answered with an independent measure of attention. In this article, we pursue Minda and Smith's (2002) agenda of independently measuring categorizers' attention but do so through different means. In particular, we conduct yet another replication of the 5–4 problem with an eyetracker.

Eyetracking and Selective Attention During Category Learning

Although little previous work has applied the method of eyetracking to category learning, a close relation between attention and eye movements has been demonstrated for many cognitive tasks. Of course, though attention can dissociate from eye gaze under certain circumstances (Posner, 1980), there is evidence that attention and eye movements are tightly coupled for all but the simplest stimuli (Deubel & Schneider, 1996). As a result, eyetracking has proven to be a useful tool in many domains of cognitive research, including reading (Ferreira & Clifton, 1986; Just & Carpenter, 1984; Mak, Vonk, & Schriefers, 2002; Rayner, 1998; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), language production (Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998) and comprehension (Frisson & Pickering, 1999), memory (Zelinsky & Murphy, 2000), scene perception (Henderson, 1999; Loftus & Mackworth, 1978), problem solving (Grant & Spivey, 2003; Hegarty & Just, 1993), skill acquisition (Haider & Frensch, 1999), and face perception (Althoff & Cohen, 1999), to name a few.

Of particular relevance to the current article is our own previous work investigating the relationship between attention and eye movements during category learning (Rehder & Hoffman, 2005). In that research, we examined how eye movements changed during the learning of the Shepard et al. (1961) category structures described earlier. As mentioned, a central finding in Shepard et al. was that the number of stimulus dimensions needed for correct classification was an important determinant of category learning difficulty. Not only did Rehder and Hoffman replicate the relative difficulty of the problem types, eyetracking data confirmed that participants learned to allocate attention in an optimal manner: By the time learning was completed, virtually all participants were fixating on only the single relevant dimension in the Type I condition and on only the two relevant dimensions in the Type II condition. Learners were fixating on all dimensions in those conditions in which all were relevant. That is, eye fixations were highly diagnostic of which dimensions participants were using to make categorization decisions by the end of learning. Moreover, we found that changes in eye fixations to only the relevant dimensions were highly correlated with reduction in classification errors. Taken together, these results suggest that eye fixations were tightly synchronized with the cognitive processes involved with learning the categories.

In this article, we use eyetracking as an independent source of evidence regarding participants' allocation of attention while learning the 5–4 category structure. Following previous investigators, we first fit both the GCM and the MPM to each participant's classification data. We then compare the attention weights that are estimated by these fits with participants' eye fixations. One possible outcome is that the GCM will suggest that substantial weight was being placed on the least diagnostic dimensions (just as in previous studies) while, at the same time, eye movements indicate that participants were never (or rarely) fixating that stimulus information. This result would be troublesome (to say the least) for the GCM, for how could classification decisions be made on the basis of information that the categorizer does not have? Indeed, this finding would confirm Minda and Smith's (2002) suspicion that the GCM's "attentional description is a formal artifice that . . . is not the best psychological description of attention strategies in the 5–4 task" (p. 282)—in essence, it would indicate that the GCM yields better fits only because it overfits the data. Conversely, participants' eye movements could reveal that most participants are, in fact, gathering information about the least diagnostic dimension, a result that would instead disconfirm the "attentional description" offered by the MPM. Although this second result would corroborate the GCM, the question would then become why category learners optimize attention for some category structures (e.g., the Shepard et al., 1961, structures) but not others (e.g., the 5–4 structure).

In addition to potentially providing an independent source of corroboration for attention allocation, the use of eyetracking with the 5–4 structure also enables us to investigate in greater detail the relation between eye movements and attention during a category learning task. One notable characteristic of the Shepard et al. (1961) category structures tested in Rehder and Hoffman (2005) is that the diagnosticity of dimensions in those structures is all or none—that is, dimensions are either (equally) diagnostic or completely nondiagnostic of category membership. In contrast, the 5–4 problem consists of dimensions that vary in diagnosticity on a continuum, from high (.77), to moderate (.66), to very low (.55). Therefore, under the assumption that learners tend to optimize selective attention, this structure allows us to explore the nature of the quantitative relation between attention and various measures related to eye movements, such as the number of fixations or the total fixation time. In particular, we begin this investigation by testing some simple hypotheses, such as whether the relation between a dimension's eye fixation time and its fitted model weight is best described as linear, as a threshold function, or as some other functional form. To this end, we examine the correspondence between fitted model weights and eye movement data for individual participants.

In summary, then, our two primary goals concern the use of eyetracking to corroborate the psychological reality of the model-determined attention weights in the 5–4 problem and to determine the relation between dimension diagnosticity and eye movements more generally. However, because the application of eyetracking to category learning is new, it provides an opportunity to address three other current issues in the psychology of category learning. Although the 5–4 problem was not specifically designed with these issues in mind, it nevertheless allows them to be addressed in at least a preliminary way.

The first additional issue we consider is how attention (as measured by eye fixations) varies as a function of the stimulus presented. Standard formulations of both prototype and exemplar models assume that, at any one time, the attention weight associated with a dimension is fixed; that is, it applies regardless of the stimulus item presented. More recently, however, Kruschke (2001) has considered models that can apply different attention weights to different exemplars. Such models can potentially account for the finding that dimensions are ignored for certain stimuli but not necessarily for others (and that attention weights may also vary as a function of the context in which a stimulus appears; Aha & Goldstone, 1990; Medin & Edelson, 1988). In the current experiment, we look for evidence of such stimulus-specific attention profiles by testing whether different test exemplars elicit different patterns of eye fixations.

The second issue concerns the influence of the perceptual salience of stimulus dimensions on eye fixations. Models of categorization have begun to distinguish the notions of *perceptual attention*, the speed and fidelity with which physical stimulus dimensions are processed, from that of *decisional attention*, the influence that perceived dimensions have on overt categorization decisions. On the one hand, research suggests that, in the special case of categories defined by a single-dimension rule, learners allocate both their perceptual and their decisional attention exclusively to that one dimension (Maddox, 2001; Maddox, Ashby, & Waldron, 2002; Maddox & Dodd, 2003). On the other hand, more generally, the two attentional systems appear to function independently in that the perceptual processing of a dimension is unaffected by its corresponding decision weight (Lamberts, 1998; Maddox et al., 2002). For example, Lamberts (1998, Experiment 2) found that although the speed with which a dimension was perceptually processed was affected by its perceptual properties, it was not influenced by the dimension's diagnosticity. In other words, perceptual attention was unaffected by decisional attention. However, whereas these previous studies often used stimuli whose features could be held in foveal vision simultaneously, the current experiment used stimuli with separable (and spatially separated) dimensions. The tendency to acquire information from these dimensions with eye movements may provide a greater opportunity for a learner's acquired knowledge of dimensions' diagnosticity to serve as a source of top-down knowledge that reduces eye fixations to (and, hence, reduces the processing of) stimulus dimensions on the basis of their perceptual salience. Accordingly, in our experiment we assess the effect of the perceptual salience on eye fixations and how that influence evolves during the course of category learning.

The third and final issue concerns the role of learning strategy. Although most past analyses of the 5–4 problem have assumed that learners memorize either exemplars or abstract prototypes, there has been a recent trend toward considering *multiple systems* theories of category learning, which presume the presence of one learning module that can discover single-dimension rules and another that memorizes exemplars (Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994; Smith, Patalano, & Jonides, 1998) or that forms a decision boundary on the basis of multiple stimulus dimensions (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & Waldron, 1999). Consistent with this proposal, Johansen and Palmeri (2002) assessed the strategies participants used while learning the 5–4 problem and found evidence for single-dimension rule testing early in learning and for exemplar-based responding later. However, evidence of early rule use was based on test blocks that were interleaved during training, which might have changed participants' learning strategy. In the current experiment, we examine participants' eye fixations early in learning for evidence of rule testing in the absence of interleaved test blocks.

Method

Participants

A total of 86 New York University undergraduates participated for class credit. We removed data from 6 participants who changed a significant proportion (4 of 9) of their responses between the training and transfer phases of the experiment, which indicates that they failed to understand the instructions for the transfer phase. We removed data from an additional 16 participants because of an excessive number of eyetracking failures (e.g., because of blinks) where the eyetracker could not locate the participants' pupil position. The remaining 64 participants were tested individually and randomly assigned to condition in equal numbers. All participants had normal vision with corrective lenses or better.

Materials

Schematic drawings of insects were used as stimuli to instantiate the 5–4 category structure. The two category prototypes are presented in Figure 1. The center of each insect was a black rectangle body. From the center of this body extended four black lines to four binary dimensions, which were head (pointed or oval), wings (wing or parachute), tail (feathered or stinger), and feet (three or five). The extent of the entire insect was an approximate square that was 12° of visual angle in width and in height. All physical features, including the body, were approximately 4° in width and height. We recorded from a single eye with the SensoMotoric Instruments (Berlin, Germany) Eyelink eyetracking system.

Design and Procedure

We used a Latin square to balance the assignment of physical dimensions to abstract dimensions so that the average diagnosticity of each physical dimension was equal across participants. This served to average over any effects on attention due to differences in the salience of any one of the dimensions. In previous experiments with these stimuli, we found that the head generally attracted more fixations, presumably either because of its low-level perceptual properties (e.g., the presence of an eye) or because of participant's preexperimental knowledge, which led them to expect that the head was likely to be important in discriminating the categories. One question we ask in this research is whether this effect changes during the course of learning.

Because the category structure was asymmetrical in that Category A contained five exemplars and B contained four, we took the additional precaution of counterbalancing the assignment of physical feature to the abstract dimension values so that, for example, half the participants saw the

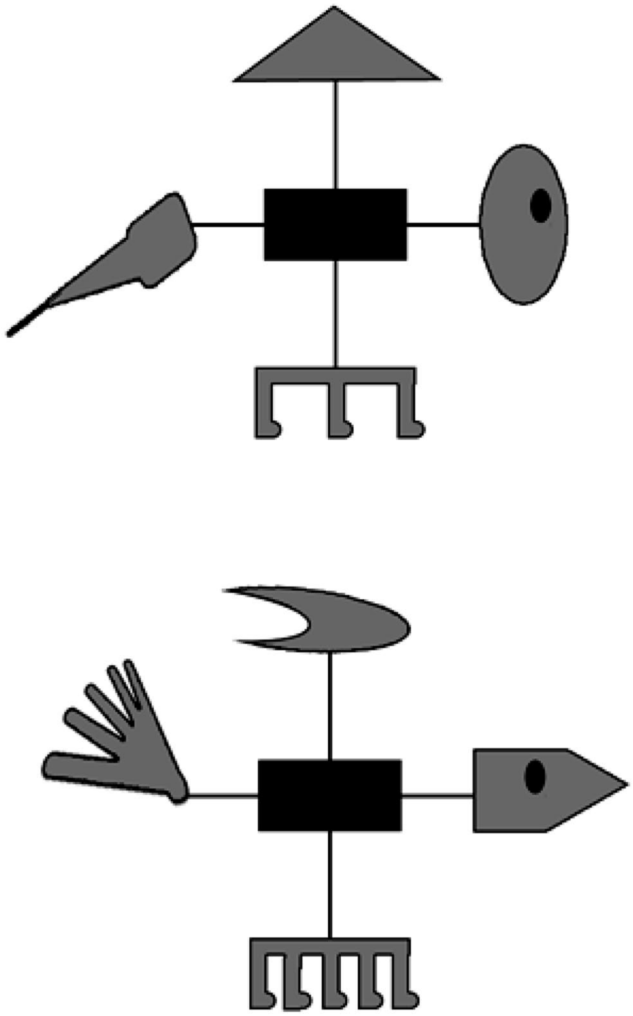

Figure 1. Stimulus examples of the two category prototypes.

feature “round head” associated with Category A and “pointy head” associated with Category B, whereas this assignment was reversed for the other half.

The rest of the experimental procedure was similar to the original 5–4 study conducted by Medin and Schaffer (1978). It consisted of a training phase and a transfer phase. In the training phase, participants were read standard category learning instructions and were told which of the insect’s features would vary from trial to trial, that they would begin the experiment by guessing which stimuli belonged to either category, and that through feedback they would eventually be able classify all nine insects. The feedback was a chime sound for a correct response and a buzzer sound for an incorrect response, which were played once to the participants before the experiment started so that they knew what to expect.

Each learning trial began with a drift correction in which the participant fixated a small circle at the center of the CRT, allowing the eyetracker to make small calibration adjustments that compensate for slight movements (drifts) of the eyetracker on the participant’s head. When a stimulus appeared on the CRT, participants responded by pressing a red or green button on a button box (assignment of color to category label was balanced). A stimulus remained on the screen for an additional 4,000 ms after feedback.

Participants classified blocks of nine training stimuli in which the stimuli were presented in random order within each block. Training ended when the participant either completed 21 blocks or completed 2 consecutive blocks without error.

During the transfer phase, participants were presented the 9 original training stimuli as well as 7 new transfer items. These 16 stimuli constitute the entire set of possible combinations of features given a four-dimensional binary category structure (see Table 1). Transfer trials were identical to those during training except that, immediately after their classification, we removed the stimulus, and no feedback was provided. In addition, after a transfer trial, participants moved the placement of a vertical bar to the left or to the right to indicate how confident they were that the stimulus belonged to the A or the B category. The confidence rating scale had 17 positions that could indicate eight levels of confidence for either the A or the B category. The bar’s starting position in the center of the screen indicated that participants were neutral regarding whether an item was an A or a B. For data analysis, confidence ratings were scaled so that a 1 indicates maximum confidence in Category A and a 0 indicates maximum confidence in Category B. The 16 stimuli were presented in random order within each of the two transfer blocks.

Eyetracking Dependent Variables

In addition to standard learning measures, such as error rates and the number of blocks to criterion, we derived several measures based on the eyetracking data. The eyetracker software yields, for each trial, a stream of fixations and their corresponding X–Y screen locations and durations. Four areas of interest (AOIs) were defined as polygons that encompassed the physical location of each of the four features on the computer screen. All fixations outside of those AOIs were discarded, as were any fixations that occurred after the participants made their classification choice. From the remaining fixations, three different types of dependent measures were derived for each stimulus dimension.

First, *fixation probability* is a binary variable encoding whether the AOI associated with a dimension was fixated at least once on a trial. When this variable is aggregated over several trials, it represents the probability that a dimension is fixated during those trials. Second, the *number of observations* approximates the number of times a dimension is fixated during a trial. It differs from number of fixations only in that two adjacent fixations to a single dimension are aggregated into a single “observation.” From this measure, we further derived *proportion of observations*, which is a dimension’s number of observations divided by the total number of observations to all four dimensions. Third, *fixation time* is the total number of milliseconds that a dimension is fixated. Because of the non-normal nature of fixation time data, all analyses and graphs used fixation times that were log transformed. When an AOI received zero fixations on a trial, we left this value untransformed, as its log was undefined. From this measure, we computed *proportion fixation time*, which is a dimension’s log fixation time divided by the total log fixation time to all dimensions.

Results

We begin with a summary of the behavioral learning data to compare our results with standard 5–4 results. We show eye-fixation data from the learning phase to demonstrate how participants’ attention profiles changed during the course of learning. We then focus our analyses on the transfer results, where the data best inform the debate between prototype and exemplar models. In particular, we examine exemplar and prototype model fits, comparing model weights with fixation data to evaluate the psychological plausibility argument posed by Smith and Minda (2000; Minda & Smith, 2002) against the exemplar model.

Training Data

We compared our training data with the standard 5–4 results. Overall, participants committed 14% errors on their last learning block. About half of the participants, 30 of 64 (47%), reached the two perfect blocks learning criterion. The remaining 34 nevertheless attained a proportion correct that was better than chance (*M* = .74, *SD* = .17; $p < .01$). These data are similar to previous 5–4 studies. For example, participants achieved an overall 18% error rate on the last block in the original Medin and Schaffer (1978) study, and 19 of 32 participants (59%) achieved errorless runs. Similarly, 36 of 96 participants (38%) achieved errorless performance in Medin and Smith (1981). Thus, the current results are consistent with previous 5–4 studies, confirming that, although perfect performance is difficult to achieve with the 5–4 structure, most participants showed general improvement.

Eye Fixations During Training

We begin our presentation of the eyetracking results by showing in Figure 2A the probability that a dimension is fixated as a function of its diagnosticity and block of training. For purposes of constructing Figure 2A, we assumed that those participants who completed training before the 21st block would have continued fixating as they had during their last 2 actual training blocks. Figure 2A illustrates that, as expected, given the counterbalanced assignment of physical dimensions, the dimension fixation probabilities were approximately equal early in learning. Over blocks, the fixation probabilities began to reflect the diagnosticities of the four dimensions. Fixations to the high dimensions increased, those to the medium dimension stayed the same, and those to the low dimension decreased. By the last 4 blocks of training, the fixation probabilities for the High 1, High 2, Medium, and Low dimensions were .77, .79, .76, and .61, respectively.

We tested the reliability of this effect by comparing the four fixation probabilities in the first four blocks with those of the last four blocks of learning in a 4 × 2 × 4 mixed design analysis of variance (ANOVA), in which dimension (High 1, High 2, medium, and low) and learning block (first four and last four) served as within-subject variables. The four assignments of diagnosticity to physical dimensions served as a between-subjects variable. The main effect of dimension was reliable, $F(3, 180) = 6.1$, $MSE = 0.06$, $p < .01$, indicating that a dimension's diagnosticity affected its fixation probability. The interaction between dimension and learning block was reliable, $F(3, 180) = 5.3$, $MSE = 0.05$, $p < .01$, confirming that the fixation probabilities changed during learning. In a separate analysis of the last four blocks of learning, planned comparisons revealed that the fixation probability for the least diagnostic dimension was less than for the medium diagnostic dimension, $F(1, 60) = 13.8$, $MSE = 0.10$, $p < .01$. The greater fixation probabilities for the two highly diagnostic dimensions compared with the medium dimension were in the expected direction but were not reliable ($F < 1$). The two high dimensions were statistically equivalent ($F < 1$).

A significant interaction was obtained between dimension and the counterbalancing factor that encoded the assignment of diagnosticity to a physical dimension, $F(9, 126) = 3.1$, $MSE = 0.01$, $p < .01$. This interaction indicates that a dimension's fixation probability was influenced by its perceptual salience. When we averaged over diagnosticity, the probabilities that participants fixated the head, tail, feet, and wings during learning were .81, .75, .72, and .61, respectively. Thus, beyond dimension diagnosticity, perceptual salience played an important role in determining fixations.

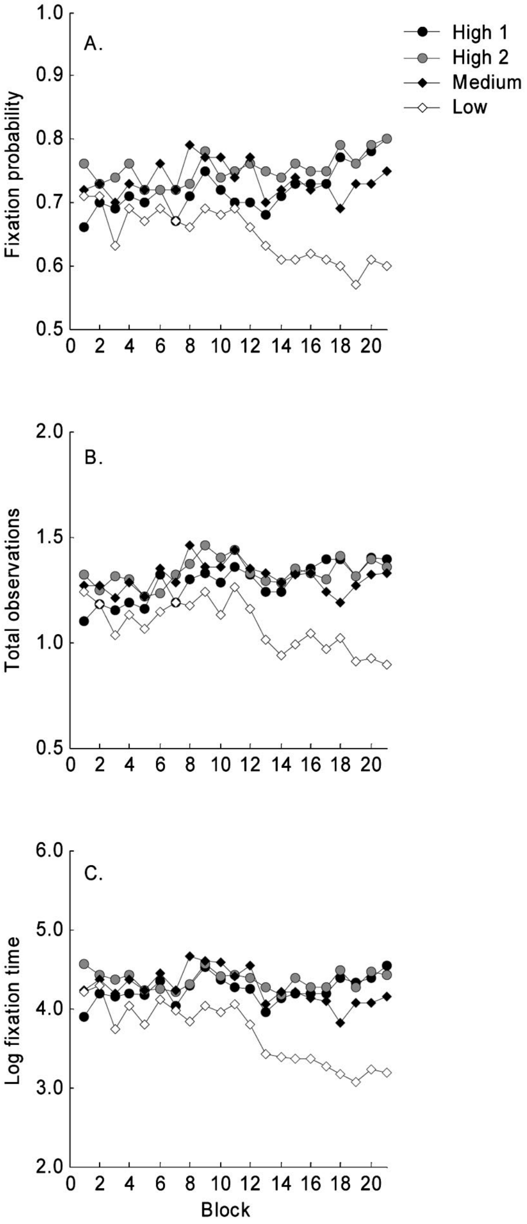

Figure 2. A. Eyetracking results during training as a function of block and dimension. A: Fixation probability. B: Total observations. C: Log fixation time.

The three-way interaction among perceptual salience, diagnosticity, and first versus last four training blocks was not reliable. That is, salience moderated the effect of diagnosticity on fixation probabilities throughout learning, and there was no change in its moderation over time.

The results shown in Figure 2A are consistent with the idea that category learning involves learning to weigh each stimulus dimension in a manner that reflects its diagnosticity. As the proportion of correct responses increased, eye fixations to the highly diagnostic dimensions increased, and those to the least diagnostic dimension decreased. The *t* tests contrasting the average fixation probabilities of the first four and last four blocks of learning showed that decreased fixations to the least diagnostic dimension and increased fixation probabilities to the first highly diagnostic dimension were reliable at $p < .05$. In addition, as expected, the medium diagnostic dimension was not fixated reliably less in the last four blocks ($t = 1$). The second highly diagnostic dimension was not fixated reliably more at the end of learning compared with the earlier blocks, but the effect was in the expected direction (.79 in the last four blocks from .74 in the first four), $t(63) = 1.4$, $p = .17$. However, if we combine the two (logically equivalent) highly diagnostic dimensions, the corresponding contrast is, in fact, reliable at $p < .05$. Thus, the pattern suggested by Figure 2A was confirmed statistically. Nevertheless, note that eye fixations to the least diagnostic dimension suggest that learners continued to allocate substantial attention to that dimension—and did so despite the fact that information from that dimension was not required for successful classification. This finding is important, because it speaks to Smith and Minda's (2000) claim that fits of the GCM to the 5–4 problem that yield substantial weights on the least diagnostic dimension may not be psychologically realistic. Because of the theoretical importance of this finding, we examined the distribution of fixation probabilities to that dimension over participants to determine whether the mean fixation probability shown in Figure 2A reflects the behavior of the typical participant or whether it arose because we averaged over participants who treated that dimension very differently. Figure 3 shows the distribution of fixation probabilities for the least diagnostic dimension in the last four blocks of learning. The figure shows that although a minority of participants fixated the least diagnostic dimension rarely (10 participants exhibited fixation probabilities to that dimension less than .20), the majority of participants fixated the least diagnostic dimension on most trials.

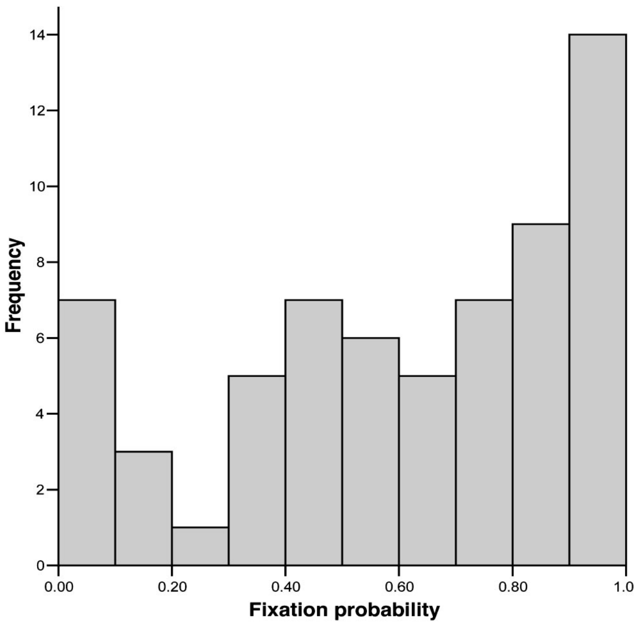

Figure 3. Histogram of fixation probability to the least diagnostic dimension.

Another possibility to consider is that although participants tended to fixate the least diagnostic dimension, those fixations were very much fewer in number or shorter in duration than fixations to the other three dimensions. Figures 2B and 2C present the training results for our two more sensitive eyetracking measures, total number of observations and log fixation time. As these figures show, the pattern of results is very similar to that for fixation probabilities in Figure 2A. Over blocks, both the total number and the duration of per trial fixations began to reflect dimension diagnosticity on average. For the last four training blocks, the average number of observations for the first high, second high, medium, and low dimensions were 1.39, 1.43, 1.37, and 0.98, respectively; the log fixations times for those dimensions were 4.65, 4.68, 4.49, and 3.44. These results indicate that, even according to these more sensitive measures, participants devoted considerable attention (as measured by eye movements) to the least diagnostic dimension.

The same 4 × 2 × 4 ANOVA that was conducted on fixation probabilities was carried out on the number of observations and log fixation times, with the same results. In both analyses, the main effect of dimension was reliable, $F(3, 180) = 7.6$, $MSE = 0.30$, for observations; $F(3, 180) = 7.5$, $MSE = 2.30$ for time, both $p\text{s} \leq .01$. In addition, the interaction with dimension and learning block was reliable, $F(3, 180) = 6.7$, $MSE = 0.13$, and $F(3, 180) = 6.2$, $MSE = 1.20$, both $p\text{s} \leq .01$, indicating that the number of observations and fixation times both changed to reflect diagnosticity during learning. For the last four blocks, the least diagnostic dimension was observed less and fixated for a shorter overall duration than the medium diagnostic dimension, $F(1, 60) = 16.4$, $MSE = 0.60$; $F(1, 60) = 17.3$, $MSE = 4.00$, $p\text{s} \leq .01$. The medium and two highly diagnostic dimensions were all statistically equivalent on both measures ($F\text{s} < 1$). In addition, significant interactions were obtained between diagnosticity and the counterbalancing factor that encoded the assignment of diagnosticity to a dimension, $F(9, 180) = 4.5$, $MSE = 0.30$, $p < .01$, and $F(9, 180) = 2.9$, $MSE = 1.80$, $p < .05$, indicating a perceptual salience effect. The average numbers of observations for the head, tail, feet, and wings during learning (averaged over diagnosticity) were 1.62, 1.32, 1.27, and 1.29, respectively; the log fixation times to those dimensions were 4.95, 4.41, 4.09, and 3.79. These results again support the finding that the head tended to have greater perceptual salience than the other three features. Finally, in neither analysis was the three-way interaction among perceptual salience, diagnosticity, and first versus last four training blocks reliable, which again indicates that the effect of perceptual salience did not change during learning. Thus, we see that the pattern of results is the same regardless of whether we examine fixation probabilities (Figure 2A), number of observations per trial (Figure 2B), or fixation times (Figure 2C).

The final aspect of the learning data we discuss is the pattern of eye fixations in the first few blocks of training. As we have mentioned, a number of current theories suggest that at the start of training learners search for low-dimension rules to distinguish the two categories. On this basis, we raised the possibility that learners would begin by fixating only one or two stimulus dimensions. In fact, we found that the average fixation probability for the four dimensions in the first four blocks was .71 (Figure 2A), a result that indicates that most dimensions were being fixated during those blocks and that corresponds to (.71 × 4 =) 2.8 out of 4 total dimensions fixated. Only 9 of 64 (14%) of our participants fixated an average of 2 or fewer dimensions early in learning. This finding is consistent with that of Rehder and Hoffman (2005), who also found that learners fixated most (2.5 out of 3) stimulus dimensions early in learning. However, rather than concluding that people were not testing single-dimension rules, Rehder and Hoffman argued that participants were pursuing learning strategies in addition to rule testing, at least one of which required information from all stimulus dimensions (e.g., exemplar memorization). We discuss this issue at greater length in the Discussion.

Transfer Data

We first compared the transfer classification results with those from previous 5–4 experiments. Figure 4 plots the average binary classifications and confidence ratings for the 16 test exemplars, along with the average classification responses from the 30 data sets reported in Smith and Minda (2000). In Figure 4, exemplars a1 through a5 are the Category A training exemplars, b1 through b4 are the Category B training exemplars, and t1 through t7 are the seven transfer items. The classification responses are the proportion of time that participants selected Category A as a response for each of the 16 test exemplars; confidence ratings reflect participants' confidence that the test exemplar was a member of Category A. As the figure shows, participants correctly selected Response A the majority of the time for the Category A training exemplars and correctly selected Response B for Category B training exemplars. For both classification decisions and confidence ratings, our data are qualitatively and quantitatively similar to the profile plot derived from averaging the response data from the 30 previous data sets. These findings indicate that our participants performed similarly to those in previous studies despite wearing an eyetracker.

As discussed, the 5–4 category structure includes two items, A1 and A2, that distinguish between prototype and exemplar models. A paired *t* test revealed that proportion correct during learning of A2 responses ($M = .76$) was reliably greater than that of A1 ($M = .71$), $t(63) = 2.0$, $p < .05$. However, it was not reliable during transfer (M = .88) for A2 and ($M = .87$) for A1 ($t < 1$); the average confidence rating of A2 ($M = .78$) was also statistically equivalent to that of A1 ($M = .77$; $t < 1$). Thus, consistent with the original Medin and Schaffer (1978) study, an A2 advantage was obtained during training but not during transfer.

A more sensitive test of whether the prototype or exemplar model provides the best account of the transfer data is to fit those models quantitatively. We fit a five-parameter version of the GCM and a four-parameter version of the MPM to each participant's transfer data. (Details regarding the fitting of these models are provided in the Appendix.) We used the sum of squared deviations (SSD) as our goodness of fit measure. Consistent with previous

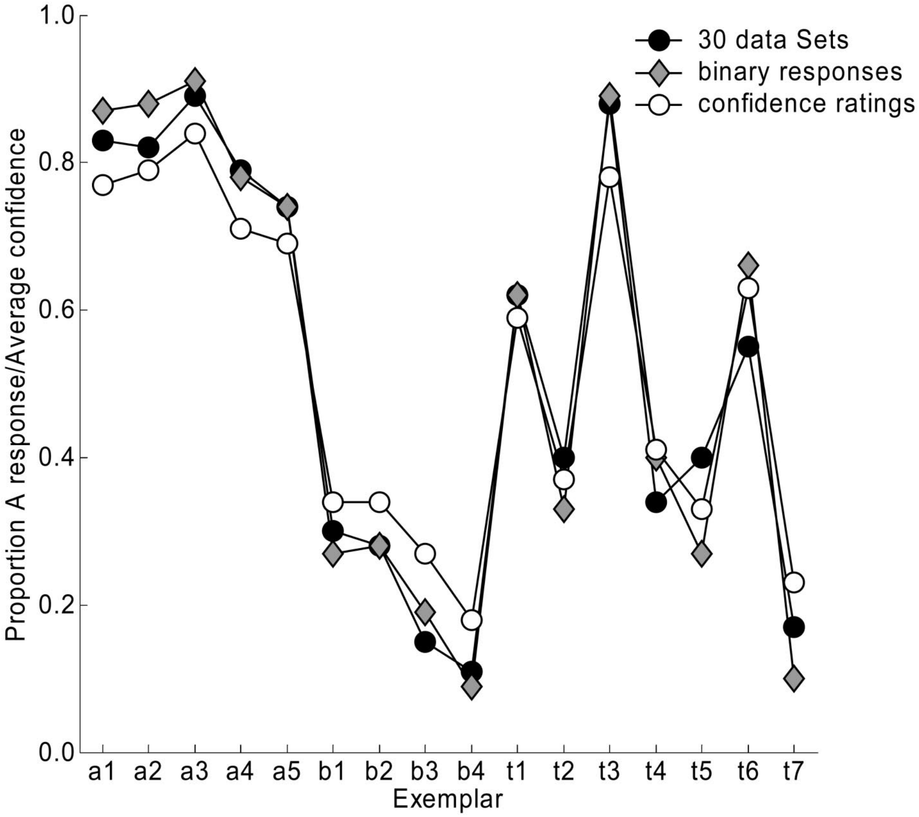

Figure 4. Response profiles for the 30 data sets reported in Smith and Minda (2000) and the binary responses and confidence ratings from the current study.

studies, we found the GCM to be quantitatively superior to the MPM. Of the 64 participants, the majority (47, or 73%) were best fit by the GCM. Overall, the average SSD for the MPM was statistically worse ($M = 0.45$, $SD = 0.39$) than for the GCM ($M = 0.35$, $SD = 0.37$), $t(63) = 2.8$, $p < .01$.

Following Minda and Smith (2002), we present in Figure 5A and 5B the model weights for the MPM and the GCM, respectively, averaged over participants. (In Figure 5, error bars depict 95% confidence intervals.) As expected, the average MPM and GCM weights look quite different. On the one hand, the MPM allocates most attention to the two high and medium dimensions the most while drastically reducing the weight on the low dimension. On the other hand, the GCM weighs the dimensions more evenly. These two attention profiles are quite similar to those found by Smith and Minda (2000) for multiple data sets. It is important to note that the MPM's low weight on the least diagnostic dimension is consistent with idea that attention gets optimized during category learning, because in fact that dimension is only weakly predictive of category membership (and can be ignored entirely). In contrast, the GCM's substantial weight on the least diagnostic dimension suggests that participants failed to optimize attention during the course of learning. It is precisely this pattern of weights that led Smith and Minda (2000) to challenge the validity of the GCM.

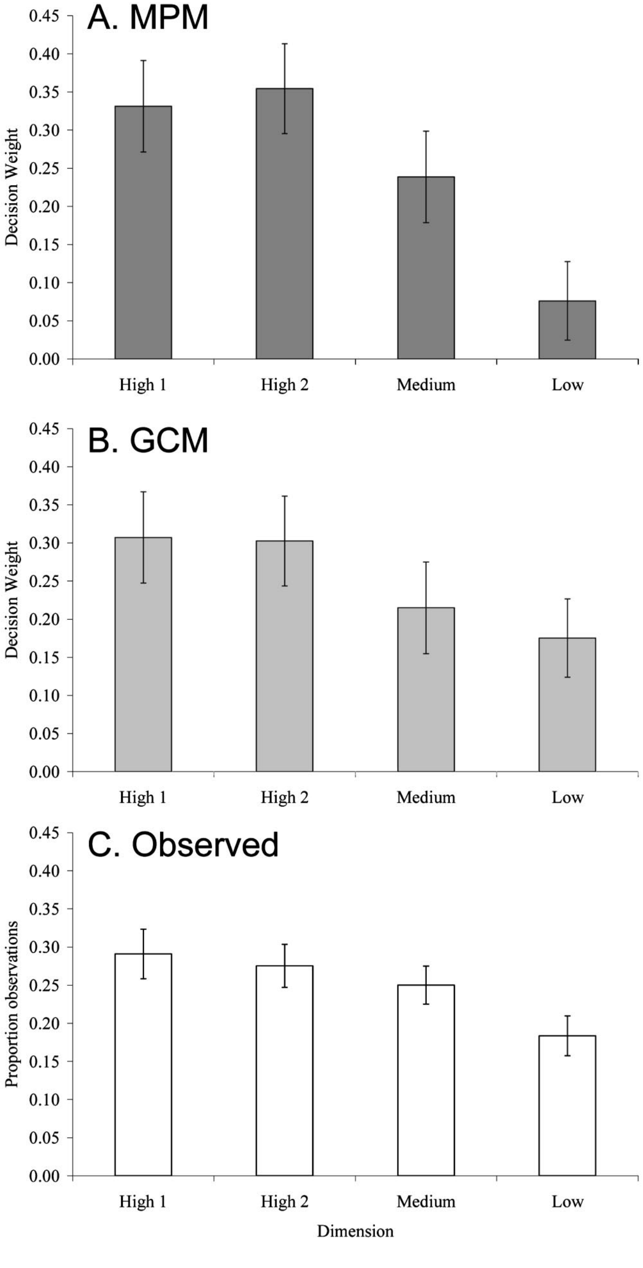

Figure 5. A: The multiplicative prototype model's (MPM's) average estimated decision weights across the four dimensions. B: The generalized context model's (GCM's) average estimated decision weights. C: Eye-tracker-measured decision weights by proportion of observations to the four dimensions.

Eye Fixations During Transfer

The central goal of the present study is to make use of eye-tracking data to corroborate the psychological reality of the attention weights yielded by model fits. Because fixation probabilities, number of observations per trial, and fixation during learning times all yielded similar results, in this section we report only the number of observations per trial and compare those observations with the theoretical model weights produced by the GCM and MPM. Note, however, that because those model weights are constrained to sum to one, they represent the relative attention allocated to each stimulus dimension. Accordingly, in this comparison we use *proportion number of observations (per trial*), which represents the relative number of fixations to a dimension (sums to one).

The proportion numbers of observations to each dimension during transfer were .29, .28, .25, and .18 for the two high, the medium, and the low dimensions, respectively. The proportions are presented graphically in Figure 5C. The figure shows that, as was the case for the learning data, participants spent a great deal of time gathering information from the least diagnostic dimension. Overall, Figure 5C shows that the proportion of observations to the four dimensions was closer to the pattern of attention weights produced by the GCM (Figure 5B) than to that produced by the MPM (Figure 5A). In particular, the MPM severely underestimated the amount of time participants fixated the least diagnostic dimension. These results indicate that the GCM is superior not only in fitting the behavioral data but also in its correspondence between attention weights and eye fixations.

The data from Figure 5C were submitted to a 4 × 16 × 4 mixed design ANOVA, in which the assignment of diagnosticity to physical dimensions served as the between-subjects variable, whereas test exemplar and proportion number of observations per dimension were within-subject variables. The main effect of dimension was reliable, $F(3, 180) = 9.5$, $MSE = 0.20$, $p < .01$. Planned comparisons revealed that the proportion of observations to the least diagnostic dimension ($M = .18$) was less than to the medium diagnostic dimension ($M = .25$), $F(1, 60) = 12.1$, $MSE = 0.40$, $p < .01$. The average proportion numbers of observations associated with the two highly diagnostic dimensions were .29 and .28; they were not statistically different from the medium dimension ($F < 1$). These data indicate that the effect of dimensions' diagnosticity was present during transfer as at the end of learning.

As during learning, there was a significant interaction between diagnosticity and the counterbalancing factor that encoded the assignment of diagnosticity to a dimension, $F(9, 180) = 2.7$, $MSE = 0.20$, $p < .01$. The proportion numbers of observations to the head, tail, feet, and wings during transfer were .31, .22, .24, and .22, respectively. Thus, salience continued to play an important role in determining the number of times a participant made an observation to a dimension.

We also tested whether an interaction was obtained between dimension and test exemplar. This interaction provides information about whether eye fixations were relatively consistent over the 16 different exemplars shown in Table 1 or whether different exemplars elicited different patterns of eye fixations, a finding that would provide support to models that potentially apply different attention weights to different exemplars (Kruschke, 2001). In fact, the interaction between diagnosticity and test item was not reliable ($F < 1$). We also conducted more focused tests asking whether eye fixations varied between certain subsets of exemplars. For instance, we asked whether the nine older items (A1–A5 and B1–B4) were fixated differently than the seven new test items (T1–T7). We reasoned that because participants repeatedly classified the old items during training, they might have settled on a particular attention profile for those items, one that might have changed when a new, unfamiliar item was presented. In fact, however, we found no reliable differences between the attention profiles for old and new items. Next, because of their theoretical importance, we tested whether A1 and A2 elicited different fixations. These items can be distinguished on the basis of the least diagnostic dimension, and we therefore speculated that additional attention might be allocated to that dimension for these items. Nevertheless, we found no reliable differences across these items either. Finally, we examined eye fixations to Category A versus Category B items, with the intuition that items may be processed differently depending on their category membership. Again, statistical tests suggested no differences between eye fixation patterns across the Category A and Category B items.

Our proportion number of observation measure provides an index of the (relative) amount of time participants were fixating a dimension over the course of an entire classification trial. However, another unique capability of the eyetracker is that it provides information about the order in which stimulus dimensions are fixated. To investigate fixation order, we divided each transfer classification trial into 50-ms bins, and in each bin we tabulated, for each dimension, whether that dimension was fixated in that bin. Aggregating these tabulations over all classification trials yielded the probability that each dimension was fixated in each bin, as shown in Figure 6. Consistent with the results shown in Figure 5C, this figure indicates that participants were more likely to fixate the two highly diagnostic dimensions than the medium dimension, which, in turn, was more likely to be fixated than the least diagnostic dimension. The new result is that dimension diagnosticity was also reflected in fixation order. Whereas the probability of fixating the two highly diagnostics dimensions reached a maximum at about 500 ms after stimulus onset, the probability of fixating the medium dimension was at a maximum at about 1,000 ms. Although the histogram for the least diagnostic dimension does not have as discernable a peak, it is clear that fixations to that dimension became relatively more probable later in the trial. Thus, we see that dimension diagnosticity was also reflected in the order in which participants fixated dimensions. Of course, Figure 6 also corroborates the claim that substantial numbers of participants were gathering information about the least diagnostic dimension, a result consistent with the GCM's theoretical model fits but not with those of the MPM.

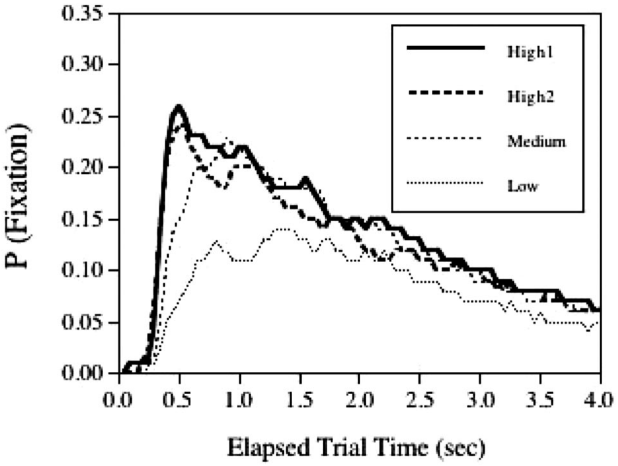

Figure 6. Changes in the probability of fixating the four dimensions from the beginning of the trial to 4 s after stimulus onset. Fixations are aggregated in 50-ms intervals.

Relating Eye Fixations and Model Weights

Having established that eye fixations corroborate the psychological reality of the GCM over that of the MPM for the 5–4 structure, we turn to our second goal, to investigate the quantitative relation between eye movements and the GCM's theoretical model weights. To this end, we performed regression analyses in which several different eyetracking measures were regressed onto the GCM model weights for each of the four stimulus dimensions. Given that the GCM weights matched the average eyetracking measures, an additional advantage of these analyses is to provide yet a stronger test of the validity of GCM model weights. If eye fixations indeed provided an independent measurement of attention during categorization, then a positive relation between eyetracking measures and model weights should be observed. That is, those participants whose GCM model fits suggest a high attention weight on a dimension should show relatively more fixations to that dimension, and those with a low attention weight on a dimension should exhibit fewer. To increase statistical power, in each regression we included three contrast-coded predictors that encoded the four ways that we assigned dimension diagnosticity (High 1, High 2,

Table 2
Partial Correlations Between Decision Weights and Eyetracking Measures

Measure	Dimension				Overall
	High 1	High 2	Medium	Low	
Fixation probability	.31*	.19	.20	.27*	.20**
No. observations	.31*	.09	.15	.31	.14*
Proportion observations	.27*	.34**	.36**	.46**	.24**
Log fixation times	.32*	.21†	.26*	.25†	.20**
Proportion log fixations	.21†	.33**	.41**	.43**	.24**

† $p < .15$. * $p < .05$. ** $p < .01$.

Medium, and Low) to the physical dimensions (head, tail, wings, or feet).[2]

The first four columns of Table 2 present the partial correlations between GCM model weights and five different eyetracking measures for each of the four dimensions. As expected, a positive correlation was obtained between a dimension's eyetracking measures and its GCM weight. Indeed, over two thirds of the statistical tests indicated a significant relation between the GCM weight and selective attention. These findings provide additional evidence that our eyetracking dependent variables were sensitive to the decision processes active in classification.

Table 2 indicates that the five eyetracking measures varied in their degree of correspondence to the GCM's model weights. To assess the relative quality of those measures, we combined the data from all four dimensions in one regression. The results are presented in the last column in Table 2. The correspondence between fixation probabilities and model weights (partial correlation of .20) was better than that for number of observations (partial correlation of .14). Additionally, there was a substantial increase in the magnitude of the correlation (to .24) by the use of proportion observations rather than the number of observations itself. This increased correspondence can be understood in terms of the theoretical interpretation of the GCM's model weights, which (because they sum to one) represent the relative weight of each dimension on categorization decisions.

Table 2 also presents the partial correlations associated with log fixation times and proportion fixation time. As was the case for number of observations, we found that the proportion log fixation measure yielded a larger correlation with model weights compared with the basic log fixation measure (.20 and .24, respectively). One might have expected that, because fixation times are based on the finest grained eyetracking measure (number of milliseconds fixating a dimension), this measure would show the strongest correspondence with model weights. In fact, however, the correlation with GCM model weights was about the same regardless of whether proportion observations or proportion fixation times were used. That is, the amount of time a learner fixated a dimension provided no additional information about the influence that dimension had on classification decisions.

We summarize the relation between our best eyetracking measure—proportion number of observations—and the GCM's model weights in Figure 7 for each of the four dimensions. In Figure 7, we have adjusted the proportion number of observations to a dimension by its perceptual salience (by controlling for the three predictors encoding physical dimension; see Footnote 2). The regression line representing the relation between model weight and salience-adjusted proportion observations is superimposed on each graph. These figures confirm the positive relation between eye fixations and model weights suggested by Table 2: When participants spent more time fixating a dimension, the GCM tended to weigh that dimension more heavily, and, conversely, when participants did not fixate a dimension, the GCM tended to weigh that dimension less.

Figure 7 suggests a linear relation between GCM weights and proportion observations. We also considered the possibility that the relation between eyetracking dependent measures and proportion observations was something other than linear by testing whether there was any sign of a quadratic or cubic relation between each eyetracking measure and model weight. All analyses failed to find such higher order relations in the data.

As mentioned, our regression analyses included contrast codes representing the physical dimension. Consistent with the above findings, the combined variance accounted for by these predictors was statistically significant in every analysis, indicating that perceptual salience influenced how people gathered information from a dimension. We also failed to find any interactions between perceptual salience and model weight, which suggests that the correspondence between selective attention and GCM weight was not moderated by the perceptual salience of a dimension.

Discussion

In this study, we used eyetracking to assess the claims of two major theories of categorization regarding how people allocate attention while learning categories. The first section below reviews how the eyetracking results speak to arguments regarding the psychological reality of the exemplar-based GCM versus the MPM. Then we attempt to reconcile our current findings of suboptimal attention allocation in the 5–4 problem with the general finding in the categorization literature showing selective attention to be optimal. We discuss more generally the relation among eye movements, selective attention, and dimension diagnosticity in category learning and then close with the following three issues:

[2] For example, one predictor coded the contrast between the head and the average of tail, feet, and wings; another coded the contrast between tail and the average of feet and wings; a third coded the contrast between feet and wings. These predictors together account for the variance in eye movements due to differences in the salience of these four dimensions.

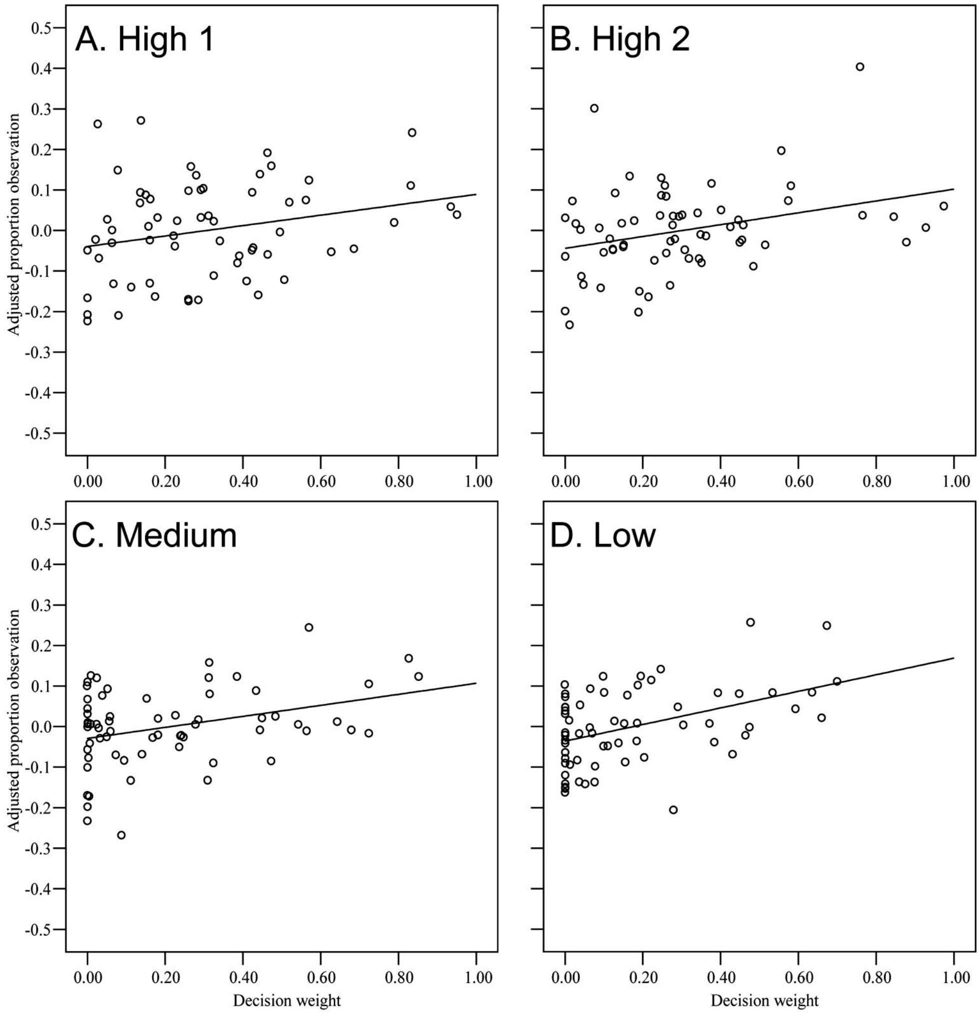

Figure 7. Generalized context model estimated decision weight plotted against proportion of observations (adjusted for the assignment of diagnosticity to physical feature). A: First high dimension. B: Second high dimension. C: Medium dimension. D: Least diagnostic dimension.

effects of perceptual salience, stimulus specific attention, and multiple learning strategies.

Selective Attention and the 5–4 Problem

Smith and Minda (2000; Minda & Smith, 2002) criticized the GCM because it typically estimates suboptimal attention weight parameters. Of the four dimensions in the 5–4 problem, only three are necessary to achieve perfect classification performance. In fact, the least diagnostic dimension should be almost completely ignored. That the GCM typically estimates substantial weight on the least diagnostic dimension, whereas the MPM estimates very little, led Smith and Minda to suggest that the MPM may be a more psychologically realistic model of learning for the 5–4 problem on the assumption that learners tend to optimize attention while learning categories. To test Smith and Minda's claim, we asked whether categorizers attend optimally in learning the 5–4 category structure, using eyetracking as another measure of selective attention. First, we found that the GCM's parameter estimates suggested that our participants placed substantial weight on the least diagnostic dimension, whereas those of the MPM suggested that the dimension received very little weight, replicating previous research. As in previous research, we found that the GCM yielded quantitatively better fits as compared with the MPM when the models were fitted to individual participant data. Our new finding is that eyetracking measures indicated that learners were, in fact, devoting a substantial number of eye movements to the least diagnostic dimension, a result consistent with the attention weights estimated by the GCM but not the MPM. In fact, the proportion of fixations to all four stimulus dimensions was remarkably close numerically to the GCM weights. These results provide corroborating evidence that the GCM is an accurate description of the underlying psychological processes in play during learning of the 5–4 problem.

The claim that the eyetracking results corroborate the GCM rests, of course, on the assumption that eye movements are in fact serving as an alternative measure of selective attention—that is, the weight that each dimension has on categorization decisions. However, it is important to ask whether there could be other reasons for the observed pattern of eye movements, factors unrelated to how stimulus information influenced classification decisions. For example, a skeptic might suggest that the relatively even allocation of eye fixations over the four stimulus dimensions arose simply because each dimension was perceptually salient. Alternatively, category learners might have fixated all dimensions because they were not only trying to classify the item but perhaps also learning which features were characteristic of each category (i.e., they were trying to learn not just features' *cue validity*, the probability of the category given the feature, but also their *category validity*, the probability of a feature given the category). Clearly, if eye fixations were measuring only features' perceptual salience or were an attempt to learn the features' prevalence within a category, those fixations would be irrelevant to arguments about the GCM's and MPM's attention weights, which are intended to reflect only the influence that a dimension has on categorization decisions.

In fact, however, there are several reasons to believe that eye fixations were not just reflecting the perceptual characteristics of our stimuli or the learning of category validities. First, we refer again to our previous study, reported in Rehder and Hoffman (2005), which used the category structures originally tested by Shepard et al. (1961). In that study, we showed that participants learned to fixate only those dimensions relevant to classification. Moreover, the restriction of eye movements to the relevant dimensions began at virtually the same time (within a few trials) as the reduction in classification errors. In other words, in that study, learners readily directed eye movements away from features irrelevant to classification and did so despite the features' perceptual salience or usefulness to learning category validities.

One difference between Rehder and Hoffman (2005) and the current study concerns the perceptual characteristics of the stimuli used. Whereas Rehder and Hoffman used relatively simple and meaningless stimulus dimensions (e.g., one dimension contrasted the letters *x* and *o*), the features of the current stimuli (schematic insects) might have been more salient and meaningful and thus more likely to attract fixations. However, in an unpublished replication of our previous study, we also used schematic insects (albeit with three dimensions instead of four) and found that participants optimized their attention by the end of learning every bit as much as they did with the less meaningful stimuli. That is, there is nothing about the perceptual characteristics of schematic insects that compels learners to fixate every feature.

A second reason for believing that eye fixations reflected decision weights is that they did reflect (if imperfectly) the underlying diagnosticities of the four stimulus dimensions. At the end of the experimental session, eyetracking measures showed increased fixations to the high dimensions, little (if any) change for the medium dimension, and a decrease for the low dimension. Thus, as in Rehder and Hoffman (2005), we found the eyetracking data to be consistent with learners selectively attending to each stimulus dimension in a manner that reflected its diagnosticity.

Finally, we found that the correspondence between eye fixations and decision weights held not only at the group level but also at the level of individual participants, as we found statistically significant correlations between individual learners' model weights and the various eyetracking measures. That is, when the GCM estimated a larger weight on a stimulus dimension, participants were more likely to fixate it, and when it estimated a smaller weight, they were less likely to fixate it. Apparently, GCM weight and eye fixations each measure a common underlying construct: the weight a stimulus dimension has on classification decisions.

Smith and Minda (2000; Minda & Smith, 2002) argued that the GCM's model fits should be held suspect because they indicate that learners adopt a suboptimal pattern of attention allocation while learning the 5–4 problem. Against this claim, we suggest that the current eyetracking results, combined with the superior fits of the GCM, provide a strong case for the GCM as a psychologically realistic description of the processes involved in learning this category structure. For better or worse, when people learn the 5–4 problem they persistently attend to a stimulus dimension that yields virtually no information about correct category membership.

Optimizing Selective Attention and the 5–4 Problem

Having established that learners indeed fail to optimize attention in the 5–4 problem, we must ask why. This result is a puzzle, because it runs counter to findings in the categorization literature demonstrating optimization. As mentioned, the classic results from Shepard et al. (1961) have been interpreted as indicating that people learn to attend to only those dimensions relevant to correct classification (an interpretation bolstered by our own previous eyetracking results). The issue confronting categorization theorists therefore concerns why categorizers optimize attention for some category structures but not for others.

One possible explanation is that, in many studies of the 5–4 problem, participants might not have been given a sufficient number of learning trials to fully optimize attention. For example, in the current study, less than half of our participants reached the two perfect blocks learning criterion. Of those who did, only 11 responded perfectly to the nine training items during the two transfer blocks, which suggests that even the majority of our so-called learners might not have fully learned the 5–4 structure. In the absence of complete learning, the failure of learners to optimize attention may be unsurprising. We do not know whether people would eventually optimize attention if training were continued for a large number (e.g., a dozen) of error-free blocks. However, we do know that even the 11 participants in our study who responded perfectly to the training items during transfer spent a great deal of time gathering information from the least diagnostic dimension. Whereas for all participants the average proportions of observations were .29, .28, .25, and .18 for the two high, the medium, and the low dimensions, respectively, for these 11 participants the proportions were .28, .29, .22, and .21, indicating that these participants fixated the least diagnostic dimension as often as (or slightly more than) the remaining participants. Moreover, the average GCM weights for these 11 participants were .31, .30, .21, and .18, indicating that the participants' overt categorization decisions were also sensitive to the least diagnostic dimension. Thus, even those participants who exhibited complete learning of the 5–4 problem showed no more sign of having optimized their attention than the rest of the participants.

Another explanation to consider is whether the value in optimizing attention for the 5–4 structure is small relative to other category structures. If optimizing attention only yields small improvements in classification, then the difference in error signal produced by optimal versus suboptimal attention profiles may not sufficiently influence a categorizer's set of decision weights. To test this conjecture, we compared the value of optimizing attention in the 5–4 problem with the value of optimizing in the Types I and II category structures from Shepard et al. (1961). Figure 8A shows the average proportion correct predicted by the GCM for the 5–4 problem with the *c* parameter set to 3 (the pattern is similar for other sensitivity values) and the attention weights set to the optimal values proposed in Lamberts (1995): .33, .33, .31, and .02 for the two high, the medium, and the low dimensions, respectively. The value of the gamma parameter (which determines the level of deterministic responding) ranged from guessing (0), to probability matching (1), to increasingly deterministic responding (greater than 1). Also shown in Figure 8A is the average proportion correct predicted by suboptimal, evenly distributed attention (all four dimension weights set to .25). Figure 8A demonstrates that there is only a slight advantage to optimizing decision weights in the 5–4 structure as compared with weighing all dimensions equally. For example, over all values of gamma, the advantage to optimizing never exceeds .02. This so-called advantage corresponds to one less error every 50 trials, or 5.6 blocks. It is questionable whether such a weak error signal could influence a categorizer's decision weights.

Contrast the small optimization advantage for the 5–4 problem with the advantage for Shepard et al.'s (1961) Problem Types I and II shown in Figures 8B and 8C, respectively. For the Type I problem, optimal attention weights were set to 1.0 for the single relevant dimension and 0.0 for the two irrelevant dimensions; for the Type II problem, they were set to 0.5 for the two relevant dimensions and 0.0 for the single irrelevant dimension. As for the 5–4 problem, the *c* parameter was set to 3. Figures 8B and 8C demonstrate a substantial advantage to optimizing attention as compared with distributing decision weights evenly. For example, when gamma equals 1, the optimization advantages for the Type I and II problems are .22 and .10, respectively; these advantages correspond to one fewer error every 4.5 trials (~2 per block) and 10 trials (~1 per block). Unlike the slight optimization advantage in the 5–4 problem (one fewer error every 50 trials), one or two fewer errors per block are likely to be sufficient to alter a categorizer's decision weights.

This difference in optimization advantage between the 5–4 structure and the Type I and II structures does not hold for all parameter values. For instance, as gamma approaches zero, performance corresponds to simple guessing and, thus, attention optimization becomes moot. Differences between optimal and suboptimal attention also disappear when gamma is very high because performance approaches ceiling. Similarly, performance is at chance when $c = 0$ and at ceiling as c approaches infinity. Nevertheless, for much of the parameter space, both Types I and II exhibit a substantial optimization advantage, whereas the 5–4 structure exhibits very little.

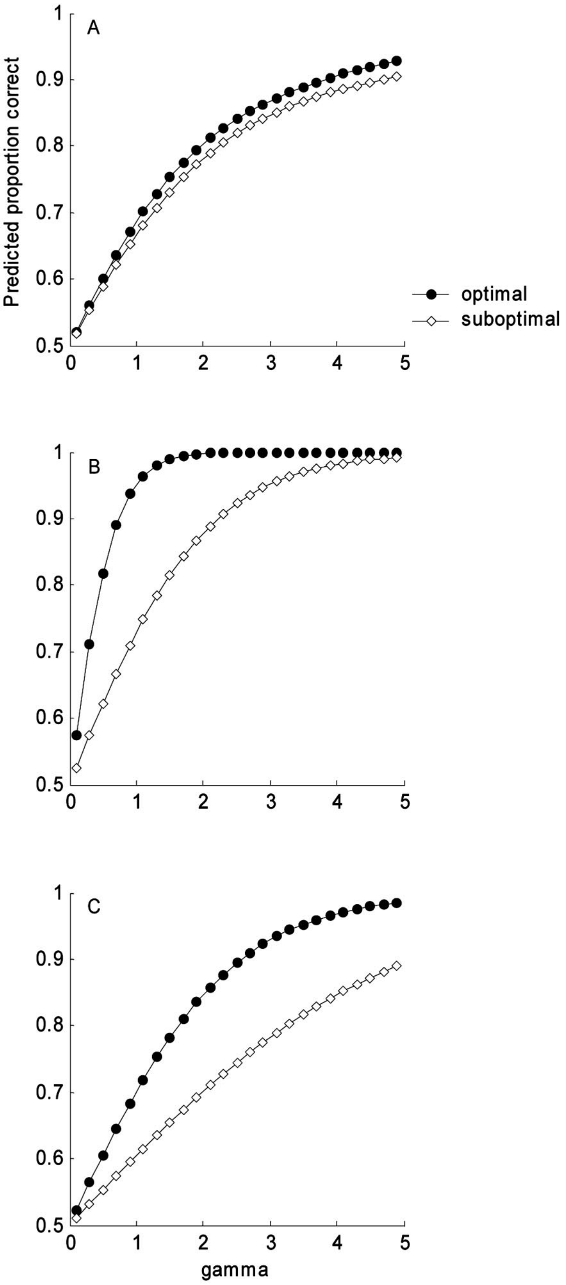

Figure 8. Generalized context model predicted proportion correct for three category structures with *c* set to 3 as a function of decision weight and gamma. A: The 5–4 structure. B: The Shepard et al. (1961) Type I structure. C: The Shepard et al. (1961) Type II structure.

Thus, our finding that fixations were consistent with an optimal attention profile in the Shepard et al. (1961) Type I and II problems but not in the 5–4 problem can be explained if we consider the differences in the advantage to optimizing attention. Whether people learning the 5–4 problem would eventually selectively attend optimally is interesting but beyond the scope of our data. We propose an *optimization advantage hypothesis* that states that people eventually optimize attention only when there is a substantial advantage in doing so (and, thus, not in the 5–4 structure).[3]

It is important to note that the advantage in optimizing attention depends not only on the category structure but also on the particular strategy that a person adopts to learn that structure. The analyses presented in Figure 8 assume an exemplar-memorization strategy. For the 5–4 structure at least, this assumption appears to be justified, as both previous research (e.g., Zaki et al., 2003) and now the current article strongly endorse the GCM as the best model for that structure. However, had learners adopted a prototype-abstraction strategy instead, we would have expected them to place much less weight on the least diagnostic dimension, because the MPM cannot solve the 5–4 problem with even attention weights.[4] The observation that a category structure's optimization advantage depends on the underlying representation is important, because some theorists have suggested that an exemplar-memorization strategy may arise only under conditions that are unlikely to hold for many real-world categories. For example, learners' propensity to adopt an exemplar-memorization strategy for the 5–4 problem has been attributed to the problem's weak *structural ratio* (the ratio of within- vs. between-categories similarity) and the fact that it has just nine exemplars and only four dimensions. When categories have a strong structural ratio and more exemplars and dimensions, there is evidence suggesting that a prototype-based learning strategy operates instead (Minda & Smith, 2001; Murphy, 2000, 2002; Smith & Minda, 2000). For these other (perhaps more ecologically valid) structures, we also predict that learners will optimize attention if there is substantial benefit to doing so, but the magnitude of that benefit will in turn depend on whether learners are in fact using prototype abstraction or the same exemplar-memorization strategy used for the 5–4 problem.

Relating Eye Fixations to Decision Weights

The second major goal of our study was to uncover the quantitative relation between eye fixations and decision weights. Understanding of this link was limited in Rehder and Hoffman (2005), because the category structures used in that study consisted of dimensions that were either equally diagnostic or completely uncorrelated with category membership. The 5–4 structure, in contrast, included dimensions whose diagnosticity varied over a continuum. Regression analyses revealed reliable linear relations between GCM weights and most eyetracking measures, with no sign of any higher order relations (e.g., quadratic, cubic). That is, to the extent that a dimension had more influence on categorization decisions, learners tended to fixate that dimension more often. Note that the strongest correlations with model weights were with proportional measures that assessed the relative number of fixations (proportion of number of observations) and relative amount of time (proportion of fixation time). This finding is consistent with the theoretical interpretation of those weights as representing the relative allocation of attention to each dimension.

One caveat concerning our claim of a linear relation between eye fixation and model weights concerns the fact that the correlations between them were substantially less than perfect, which raises the possibility that noise in our data might be obscuring a more subtle relation between these two variables. It is worth enumerating the sources of noise that may be contributing to the variance in this relation. The eyetracking measures are noisy because people blink or stare at the screen (perhaps inadvertently at a dimension). If a fixation occurs near the border between two dimensions, it may be counted as a fixation to the wrong dimension. There is also inherent noise in the eyetracker itself, because it is accurate from ~0.5° to ~1° of visual angle. Conversely, it is important to recognize that the GCM model weights themselves are imperfect indicators of underlying decision weights. For example, there is noise in participants' responses (e.g., they may hit the wrong button). This problem might have been exacerbated in the current study because the GCM model weights were estimated on the basis of a relatively small number of transfer classification trials (32). The theoretical analyses presented in the previous section showed that the GCM's 5–4 predictions were relatively insensitive to model weights, indicating that small amounts of noise in classification behavior can yield large changes in the weights (also see Lamberts, 1998). Finally, it is possible that a participant's representation of the category shifted during the transfer phase. Thus, although our analyses support a linear relation between eye fixations and model weights, it should be acknowledged that noise in both eye movements and classification choices might be obscuring a more complex relation between these two variables.

Eye Fixations and Perceptual Salience

Another goal of our study was to investigate the effects of perceptual salience on eye movements. On the basis of our previous experience with schematic insects as stimuli, we expected and found that one dimension of those stimuli (the head) would attract more eye fixations and observations compared with the other dimensions. However, the new question we asked concerned how this effect evolved during the course of learning. Previous research has suggested that when correct classification requires information from multiple dimensions to be combined, changes in decision weights alter neither the speed (Lamberts, 1998) nor the fidelity (Maddox et al., 2002) with which those dimensions are perceptually processed. However, whereas these previous results were obtained with stimuli whose features could all be held in foveal vision, the current stimuli consisted of spatially separated dimensions requiring eye movements, a condition we thought might afford our participants a greater chance of overcoming the effects of perceptual salience during learning. In fact, however, we found that participants were allocating a greater proportion of their fixations to the insect's head at the end of the experimental session every bit as much as they were at the beginning. Moreover, this effect did not interact with the diagnosticity of the head dimension. Treating eye fixations as an alternative index of dimen-

[3] One interesting question is whether the failure to optimize attention for the 5–4 structure is predicted by the exemplar-based connectionist model ALCOVE (Kruschke, 1992), which, unlike the GCM itself, specifies the mechanism by which attention weights change. Because ALCOVE requires error feedback to change attention weights and because high levels of performance can be obtained with evenly distributed (i.e., suboptimal) attention weights in the 5–4 problem, it may well be that ALCOVE, like our human participants, would solve the 5–4 problem by associating exemplars to category labels without directing attention away from the least diagnostic dimension.

[4] For example, when the four attention weights are each .25, the MPM cannot classify A2, B1, and B2 above chance, because each of these items contains two features from Category A and two from Category B (and, thus, the evidence in favor of membership is equal for the two categories).

sions' perceptual processing, we found, as in Lamberts (1998) and Maddox (2002), that such processing did not change during category learning.

There is an important exception to this generalization that perceptual processing is unaffected by learning, however. Maddox et al. (2002; also see Maddox & Dodd, 2003) found that categories that could be distinguished on the basis of a single dimension led to decreased processing of the irrelevant dimension (i.e., perceptual noise on that dimension increased). Analogously, our eye-tracking data in Rehder and Hoffman (2005), in which learners ceased fixating irrelevant dimensions entirely, can be interpreted as an extreme example of reduced (to zero) processing of irrelevant information. However, apparently, when a dimension is at all relevant to classification, the relative amount of perceptual processing it consumes remains unchanged even after substantial learning.

We have not addressed the question of why the head of the schematic insects attracted additional eye movements. One possibility is that preattentive processes operating on the head's low-level perceptual properties (its oval or bullet shape or the presence of an eye) captured attention and obligated eye movements to that screen location. Another is that learners' preexperimental semantic knowledge led them to expect that the head was likely to be important in learning the category discrimination and that this bias led them to persist in fixating the head even after a substantial number of learning blocks. However, as already mentioned, in our lab we have trained undergraduates to distinguish insects on the basis of a single dimension (i.e., the Type I Shepard et al., 1961, category structure) and found that learners were able to direct eye movements away from the head when it was irrelevant to classification. Another possibility is that the oval- and bullet-shaped heads (and their eye) required relatively more resources to process (and thus were relatively difficult to distinguish). Although such questions are beyond the theoretical scope of our article, we note in passing that eye-fixation data can potentially distinguish among some of these possibilities by, for example, measuring whether a dimension tends to be fixated early in an experimental trial (indicating that it captures attention) or whether it only consumes a greater amount of time overall (indicating that it requires greater perceptual processing).

Stimulus-Specific Fixations

In this article, we have demonstrated that an independent measure of attention provided by eyetracking data has the potential to resolve whether people always allocate attention optimally during category learning and how the effects of perceptual salience change during learning. However, we also identified two other outstanding questions that such data can shed light on. One key assumption of both exemplar and prototype models is that attention weights are invariant with respect to the particular stimulus being classified. Note that this assumption was made out of necessity, because heretofore researchers could only estimate attention weights by fitting theoretical models to an entire set of classification responses. In contrast, eyetracking data have the potential to assess how the allocation of attention varies from trial to trial as a function of the particular stimulus being presented. Contra this notion of stimulus-specific attention allocation (SSA), we in fact found no evidence that eye fixations varied as a function of which of the 16 transfer items was being classified. However, as with any statistical test, our null result might have been due to a lack of sufficient statistical power. To circumvent this problem, we conducted other tests: old versus new items, Item A1 versus Item A2, and Category A versus Category B items. However, these more focused tests also yielded null results, and we concluded therefore that our participants probably did not use SSA.

Aside from possibly insufficient statistical power, we tried to gain insight into why we did not find SSA in our study. According to Kruschke (2001), SSA amounts to a computational solution for generating nonlinear mappings between particular cue configurations and attention weights in a connectionist system. Roughly speaking, it allows the system to learn when to ignore or attend to particular dimensions. One way to do this, of course, is to have exemplar-mediated decision weights, which allow contextual cues to play an important role. For example, we know that shape cues are generally unimportant in classifying liquids (e.g., liquid shape depends only on the container), yet in some contexts, the shape of a liquid can be predictive (e.g., soda-bottle-shaped liquid is likely to be beer or soda). In fact, Kruschke argued that the psychological plausibility of SSA is based on context specificity found in Aha and Goldstone (1990). Evidence for context effects comes from a variety of areas, including the perception of odors as pleasant or unpleasant (Herz, 2003), function learning of nonlinear functions (Lewandowsky, Kalish, & Ngang, 2002), and categorization of stimuli with nonlinear decision rules (Yang & Lewandowsky, 2003).

There are therefore two crucial characteristics of the 5–4 category structure that might have prevented us from finding SSA. First, the 5–4 category structure does not require a nonlinear association of cues to attention weights. That is, a single set of attention weights is sufficient to achieve perfect classification. This is unlike the category structures used, for example, in Aha and Goldstone (1990), where one set of attention weights would not accurately classify half of the items. Second, the 5–4 category structure does not contain any additional contextual cues that could possibly help participants apply different attention weights for different cue configurations. For example, in Yang and Lewandowsky (2003), contextual cues were correlated with dimension values but not with the category label itself. These contextual cues were shown to help participants form nonlinear associations between predictive cues and attention weights. It is possible that, had we provided either contextual cues or a category structure that necessitated nonlinear mappings of cues to attention weights, we would have found SSA. Because we provided neither, it is perhaps no surprise that SSA was not revealed in our fixation data.

Learning Strategies

Finally, one important question for categorization research is the relative role of similarity-based processes (e.g., exemplar memorization) and rule- or hypothesis-testing accounts. Many prominent accounts of rule-based category learning have assumed that learners begin by testing simple single-dimension rules and only consider more complex multidimension decision rules after they find no successful single-dimension rule (Bourne, 1970; Bruner, Goodnow, & Austin, 1956; Nosofksy et al., 1994). The most straightforward prediction that such accounts make for eye movements is that category learners will fixate only a single stimulus dimension early in learning, during the one-dimension rule-testing phase. In contrast, we found that, at the beginning of the experiment, our 5–4 learners fixated most (2.8 out of 4) stimulus dimensions. This result is similar to that reported in Rehder and Hoffman (2005), who also found that learners fixated most stimulus dimensions in the first few trials of learning (2.5 out of 3).

Superficially, at least, these results challenge the view that learners test simple single-dimension rules early in learning. Specifically, the findings appear to contrast with those of Johansen and Palmeri (2002), who found evidence for early rule testing on the basis of test blocks that were interleaved with the training blocks. It is conceivable that such a procedure might have artificially induced learners to form concrete hypotheses (e.g., single-dimension rules) so that they would have some basis for responding to the test blocks.

Against this interpretation, however, Rehder and Hoffman (2005) found a number of other sources of evidence that strongly implicated hypothesis testing. For example, in that study, we examined individual participant data from the condition (Type I) in which the two categories could be discriminated on the basis of a single-dimension rule. We found that a majority of participants exhibited a shift in error rates from about .50 (random responding) to .00 in just one or two trials, a result that strongly suggests that learners (suddenly) discovered the correct rule. Moreover, changes in eye movements were highly correlated with the changes in error rate. On the basis of these results, Rehder and Hoffman concluded that, early in learning, participants were in fact pursuing multiple learning strategies in parallel (rule testing, exemplar memorization, search for meaningful relations among features, etc.), which entailed that they examine all stimulus dimensions but which also enabled them to discover simple categorization rules when such rules were present. We see a similar pattern of results when we combine the 5–4 results from Johansen and Palmeri (2002) and those of the current study: Early in learning, the behavioral data implicate rule testing, whereas eyetracking data implicate the fixation of multiple stimulus dimensions. Following Rehder and Hoffman, we suggest that such results are consistent with multiple systems theories of category learning, which presume the presence of one learning module that can discover single-dimension rules and another that memorizes exemplars (Erickson & Kruschke, 1998; Nosofksy et al., 1994; Smith et al., 1998) or that forms a decision boundary on the basis of multiple stimulus dimensions (Ashby et al., 1998; Ashby & Waldron, 1999).

Of course, learning strategies might vary not only within a single individual but also among individuals; that is, different people might apply different strategies to learn the same classification problem. For example, in the current study, we found that although the GCM yielded better fits to a large majority of participants, a substantial minority (27%) were in fact fit better by the MPM, a result that is at least suggestive of the possibility that those individuals learned by abstracting prototypes. In our future research, we plan to use eyetracking as evidence for the existence of different learning strategies. The current experiment was not designed for this purpose, because participants were presented with stimuli that were not equated for their perceptual salience (e.g., for some participants, the more salient head dimension was also highly diagnostic; for others, it was the least diagnostic dimension). However, Rehder and Hoffman (2005) found at least preliminary evidence for different learning strategies: Whereas the vast majority of participants fixated all dimensions at the beginning of learning and only relevant ones at the end, a few never fixated more than one dimension (individuals we referred to as *explicit rule testers*), and a few never stopped fixating irrelevant dimensions (individuals we referred to as *explicit memorizers*).

However, it is also important to recognize that the mere fact that a minority of participants were fitted better by an alternative model (in this case, the MPM) is no guarantee that it is the correct description of those individuals, because the existence of noise in classification responses makes it inevitable that an alternative model will sometimes produce the better fit. To illustrate this fact, we performed a power analysis to see how often the MPM fitted best when the GCM was the true underlying model. We first used the GCM to generate 2,000 simulated categorizers with distributions of sensitivity and decision weights that mirrored the distributions derived from our human data. Gamma ranged across the values 1, 2, 3, 4, and 5. For each categorizer, the GCM generated 16 choice probabilities for each of the 5–4 structure's training and transfer items shown in Table 1. Each of those probabilities was then adjusted by a random amount drawn from the distribution $N(0, \sigma)$, where σ was either .0, .1, .3, or .5. (Adjustments that yielded a probability greater than 1 or less than 0 were converted to 1 and 0, respectively.) The GCM and the MPM were then fitted to each of these 2,000 simulated categorizers. In fact, at noise levels of .0, .1, .3, and .5., the MPM yielded a smaller SSD for 0%, 11%, 16%, and 29% of those simulated categorizers, respectively.

This result indicates that the presence of a minority of participants who were fit best by the MPM is not by itself sufficient evidence for the presence of more than one learning strategy. We suggest that an important area of future research is development of the model-comparison techniques necessary to distinguish between the presence of quantitatively different learning strategies versus response noise, an endeavor for which eyetracking data may well play an especially useful role.

Conclusion

In this research, we used eyetracking technology as an additional measure of selective attention in category learning. There are two primary findings. The first is that the decision weights measured by the eyetracker were consistent with the exemplar-based GCM but not the prototype-based MPM. This finding corroborates the general superiority of the GCM in fitting learners' behavior in the 5–4 category structure. It is notable that both the GCM and the eyetracker showed participants to be using suboptimal decision weights in the 5–4 problem. We proposed an optimization advantage hypothesis as an explanation for why learners optimize decision weights for some category structures but not for others. The second finding is of an approximately linear relation between eye fixations and a dimension's decision weight. These results underscore the usefulness of eyetracking as a new tool in the study of category learning. Because it provides an independent measure of the underlying decision weights used in classification, we expect that this methodology will prove indispensable for testing current and future models of category learning.

References

Aha, D. W., & Goldstone, R. (1990, July). *Learning attribute relevance in context in instance-based learning algorithms.* Paper presented at the meeting of the Cognitive Science Society, Hillsdale, NJ.

Althoff, R. R., & Cohen, D. (1999). Eye-movement-based memory effect: A reprocessing effect in face perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 997–1010.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105,* 442–481.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 33–53.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review, 6,* 363–378.

Bourne, L. E. (1970). Knowing and using concepts. *Psychological Review, 77,* 546–556.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New York: Wiley.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36,* 1827–1837.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127,* 107–140.

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory & Language, 25,* 348–368.

Frisson, S., & Pickering, M. J. (1999). The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1366–1383.

Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: Guiding attention guides thoughts. *Psychological Science, 13,* 462–466.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11,* 274–279.

Haider, H., & Frensch, P. A. (1999). Eye movement during skill acquisition: More evidence for the information-reduction hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 172–190.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18,* 441–461.

Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1084–1102.

Henderson, J. M. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 210–228.

Herz, R. S. (2003). The effect of verbal context on olfactory perception. *Journal of Experimental Psychology: General, 132,* 595–606.

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology, 45,* 482–553.

Just, M. A., & Carpenter, P. A. (1984). Using eye fixations to study reading comprehension. In D. E. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research* (pp. 151–182). Hillsdale, NJ: Erlbaum.

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

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.

Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General, 124,* 161–180.

Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 695–711.

Lewandowsky, S., Kalish, M., & Ngang, S. K. (2002). Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General, 131,* 163–193.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 565–572.

Maddox, W. T. (2001). Separating perceptual processes from decisional processes in identification and categorization. *Perception & Psychophysics, 63,* 1183–1200.

Maddox, W. T., Ashby, F. G., & Waldron, E. M. (2002). Multiple attention systems in perceptual categorizations. *Memory & Cognition, 30,* 325–339.

Maddox, W., & Dodd, J. L. (2003). Separating perceptual and decisional attention processes in the identification and categorization of integral-dimension stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 467–480.

Mak, W. M., Vonk, W., & Shriefers, H. (2002). The influence of animacy on relative clause processing. *Journal of Memory & Language, 47,* 50–68.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 128–148.

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.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 241–253.

Meyer, A. S., Sleiderink, A., & Levelt, W. J. M. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognitive Psychology, 66,* B25–B33.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 275–292.

Murphy, G. L. (2000). Explanatory concepts. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 361–392). Cambridge, MA: MIT Press.

Murphy, G. L. (2002). *The big book of concepts.* Cambridge, MA: MIT Press.

Nosofsky, R. M. (1984). Choice, similarity and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114.

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

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 924–940.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32,* 3–25.

Rayner, K. (1998). Eye movements in reading and information processing: Twenty years of research. *Psychological Bulletin, 124,* 372–422.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology.*

Rosch, E. H., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75*(Whole No. 517), 42.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1411–1436.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition, 65,* 167–196.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995, June 16). Integration of visual and linguistic information in spoken language comprehension. *Science, 268,* 1632–1634.

Yang, L. X., & Lewandowsky, S. (2003). Context-gated knowledge partitioning in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 663–679.

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1160–1173.

Zelinsky, G., & Murphy, G. L. (2000). Synchronizing visual and language processing: An effect of object name length on eye movements. *Psychological Science, 11,* 125–131.

Appendix

Formal Modeling Procedures

In the sections that follow, we describe the generalized context model (GCM), the multiplicative prototype model (MPM), and parameter estimation procedures.

The GCM

The exemplar-based GCM represents categories by storing exemplars in memory. It classifies items by first calculating the weighted distance between to-be-classified item *i* and an exemplar stored in memory, *j*, by

$$d_{ij} = \Sigma w_k \cdot |x_{ik} - x_{jk}| \ , \tag{1}$$

where x_{ik} takes the value 0 or 1 depending on dimension *k*. Differences among dimensions are weighed by a dimensional decision weight (*attention weight*), w_k ($0 \leq w_k \leq 1$, $\Sigma w_k = 1$), and then transformed by the following exponential decay function:

$$S_{ij} = e^{-c \cdot dij}. \tag{2}$$

Equation 2 produces a similarity gradient depending on the sensitivity parameter *c*. Higher values of *c* increase gradient steepness, whereby items must be increasingly similar to a stored exemplar to be classified consistently into a category. Finally, similarities are inputted into the choice rule by

$$P(R_A|i) = \frac{S_{iPA}}{S_{iPA} + S_{iPB}}, \tag{3}$$

which states that the probability of a Category A response *R* given item *i* is the ratio of the similarities between that item and all stored exemplars *j* in Category A versus those in Categories A and B. Similarities are raised to a gamma parameter specifying the level of deterministic responding.

The MPM

The MPM is similar to the GCM in that items are classified via distances in a multidimensional space. However, distances in this model are calculated between to-be-classified items and a prototype rather than with previously classified exemplars,

$$d_{iP} = \Sigma w_k \cdot |x_{ik} - P_k| \ , \tag{4}$$

where P_k is the category prototype. Otherwise, the distance calculation functions much like Equation 1. As with the GCM's Equation 2, distance is transformed into similarity by the exponential decay function by

$$S_{iP} = e^{-c \cdot diP}, \tag{5}$$

where *c* represents the participant's sensitivity to similarity. However, with the MPM, the *c* parameter has a dual function because it also represents an individual's level of deterministic responding. Recall that with the GCM, deterministic responding was represented separately by the gamma parameter in Equation 3. Although a gamma parameter may also be added to the MPM in the following equation (Equation 6), one cannot estimate its value independently of the sensitivity parameter. In fact, versions of the MPM with and without the gamma parameter are mathematically identical (see Nosofsky & Zaki, 2002, p. 926, for a complete proof of this fact). Similarity to the Category A and B prototype is then entered into the choice equation given by

$$P(R_A|i) = \frac{S_{iPA}}{S_{iPA} + S_{iPB}}, \tag{6}$$

which functions much like Equation 3.

Model Fitting

Figure 4 shows that the average confidence ratings and binary responses were very similar, indicating that inferences about participants' strategies are likely to be equally valid with either measure. Because the binary response data were too noisy for model fitting at the individual level, modeling required the use of confidence ratings exclusively. In using confidence ratings, we assumed that participants could, in principle, provide finer data than the standard "A" or "B" response. The similarity found between average confidence ratings and binary responses supports this assumption.

The GCM and MPM were fitted to each individual participant's average confidence rating for the 16 test exemplars. The GCM fits involved five free parameters, and the MPM fits involved four. Parameter values were chosen that minimized sum of squared deviations.

Received November 9, 2004
Revision received March 8, 2005
Accepted March 15, 2005 ■