Journal of Experimental Psychology:
Human Perception and Performance
2009, Vol. 35, No. 2, 530–550

0096-1523/09/$12.00 DOI: 10.1037/0096-1523.35.2.530

Beyond Nonutilization: Irrelevant Cues Can Gate Learning in Probabilistic Categorization

Daniel R. Little and Stephan Lewandowsky
University of Western Australia

In probabilistic categorization, also known as multiple cue probability learning (MCPL), people learn to predict a discrete outcome on the basis of imperfectly valid cues. In MCPL, normatively irrelevant cues are usually ignored, which stands in apparent conflict with recent research in deterministic categorization that has shown that people sometimes use irrelevant cues to gate access to partial knowledge encapsulated in independent partitions. The authors report 2 experiments that sought support for the existence of such knowledge partitioning in probabilistic categorization. The results indicate that, as in other areas of concept acquisition (such as function learning and deterministic categorization), a significant proportion of participants partitioned their knowledge on the basis of an irrelevant cue. The authors show by computational modeling that knowledge partitioning cannot be accommodated by 2 exemplar models (Generalized Context Model and Rapid Attention Shifts 'N Learning), whereas a rule-based model (General Recognition Theory) can capture partitioned performance. The authors conclude by pointing to the necessity of a mixture-of-experts approach to capture performance in MCPL and by identifying reduction of complexity as a possible explanation for partitioning.

Keywords: probabilistic categorization, knowledge partitioning, exemplars, rules

In multiple cue probability learning (MCPL), people learn to predict a discrete outcome on the basis of one or more cues of varying but imperfect validity. For example, a cue might be associated with Outcome A on 75% of all trials, whereas Outcome B occurs on the remaining 25%. Probabilistic relationships of this type are common in a number of real-world decision-making tasks, ranging from psychological or medical diagnosis to predicting the weather; accordingly, research into MCPL has frequently been considered a tractable arena for studying real-world decisions (Estes, 1976; Kruschke & Johansen, 1999). Here, we focus on situations in which each response is immediately followed by corrective feedback, in which case MCPL becomes indistinguishable from category learning with a probabilistic assignment of stimuli to categories (Ashby & Gott, 1988; Kalish & Kruschke, 1997; McKinley & Nosofsky, 1995; Nosofsky & Stanton, 2005; Ratcliff, Van Zandt, & McKoon, 1999).

The fact that the mapping between responses (e.g., "A" or "B") and feedback (e.g., "correct" or "wrong") is probabilistic rather than deterministic implies that perfect performance is unattainable. People therefore typically learn by *probability matching*—that is, they learn to assign an outcome to each stimulus with a probability that matches its actual probability of occurrence (Friedman & Massaro, 1998; Myers, 1976; Shanks, Tunney, & McCarthy, 2002; Vulkan, 2000) or to mimic probability matching by deterministically responding in the presence of perceptual noise (Ashby & Gott, 1988; Ashby & Lee, 1991).[1] In consequence, MCPL and probabilistic categorization can be differentiated from conventional deterministic categorization on a number of empirical and theoretical dimensions.

Concerning empirical differences, people find probabilistic tasks more difficult to learn than their deterministic counterparts because identical cues are associated with opposing outcomes, and performance error cannot be eliminated (Juslin, Olsson, & Olsson, 2003; Mehta & Williams, 2002; Yamauchi & Markman, 2000; Young, Wasserman, Johnson, & Jones, 2000). Likewise, the imperfect relationship between cues and outcomes impedes the transfer of learned rules and associations in probabilistic tasks (Mehta & Williams, 2002). Finally, there is some evidence that different processes underlie performance in probabilistic and deterministic categorization (see, e.g., Rouder & Ratcliff, 2004). Broadly speaking, in deterministic tasks, exemplar representations frequently capture performance better than a rule-based approach, whereas the reverse is often true in probabilistic categorization (Juslin et al.,

Daniel R. Little and Stephan Lewandowsky, School of Psychology, University of Western Australia, Crawley, Australia.

Preparation of this article was facilitated by several Discovery grants from the Australian Research Council and an Australian professorial fellowship to Stephan Lewandowsky. Daniel R. Little was supported by a Jean Rogerson postgraduate scholarship. We thank John Kruschke for provision of his RASHNL code.

Correspondence concerning this article should be addressed to Daniel R. Little, School of Psychology (M304), University of Western Australia, Crawley, W.A. 6009, Australia. E-mail: daniel.r.little@gmail.com; URL: www.cogsciwa.com

[1] The other notable response pattern is known as *maximizing* and involves deterministically responding with the outcome that has the higher probability. Thus, when maximizing, people always choose the outcome whose actual probability exceeds .5 and never the outcome whose probability is below .5. Unlike probability matching, maximizing optimizes performance to the extent possible. In tasks in which feedback immediately follows responding, probability matching is known to be more prevalent than maximizing (Erev & Barron, 2005).

2003). However, this distinction is far from clear, and a variety of factors have been nominated as candidate explanations for the switch from one type of representation to another. Among those candidates are the confusability of the stimuli (Rouder & Ratcliff, 2004), the quality of the feedback (Juslin et al., 2003), and the shape of the optimal classification boundary (McKinley & Nosofsky, 1995).

At a theoretical level, existing approaches to MCPL differ from those in deterministic settings in at least two ways. First, the inevitable persistence of error that arises from probabilistic reinforcement requires modification to conventional learning mechanisms. For example, in the RASHNL (Rapid Attention Shifts 'N Learning; Kruschke & Johansen, 1999) model, error-driven learning is gradually attenuated, thus ultimately leading to discounting of error and, by implication, the stable utilization of imperfectly valid cues. Second, all existing theories of probabilistic categorization have assumed homogeneous representations—that is, representations that are invariant across different test situations and identical for all stimuli. In the case of RASHNL, representations are homogeneous because all encountered stimuli are equally and uniformly represented as exemplars. The same homogeneity applies to other exemplar models that have been applied to probabilistic categorization (e.g., generalized context model [GCM]; McKinley & Nosofsky, 1995) as well as to rule-based models (e.g., general recognition theory [GRT]; Ashby & Gott, 1988) that postulate that people divide a dimension of relevant cues with a rule that applies to all stimuli equally.

The homogeneity assumption underlying probabilistic theories is at odds with recent developments in deterministic settings and is being critically reevaluated in this article. To foreshadow briefly, we next review some of the evidence for heterogeneous representations in deterministic settings, with particular emphasis on the *knowledge partitioning* (KP) framework. We then present two experiments that confirm the existence of KP in MCPL, which constitutes a counterintuitive outcome in light of previous related results. Next, we show by computational modeling that partitioning cannot be accommodated by existing exemplar models. Instead, the data and the modeling point toward the need for development of a "mixture-of-experts" model of MCPL.

In deterministic categorization, much recent evidence has pointed to the existence of heterogeneous representations (Erickson & Kruschke, 1998, 2002; Jones, Maddox, & Love, 2006; Lewandowsky, Roberts, & Yang, 2006; Love, Medin, & Gureckis, 2004; Yang & Lewandowsky, 2003, 2004). The common thread underlying all those findings is that different representations (e.g., rules vs. exemplar memory) drive responding to different stimuli (e.g., Erickson & Kruschke, 1998) or that the same stimulus elicits different subcomponents of knowledge in different circumstances (e.g., Yang & Lewandowsky, 2003, 2004). Here, we are concerned primarily with the latter finding, known as KP.

The KP framework posits that knowledge, such as the representations used in categorization, may be fractionated into independent "parcels" that are used selectively and without reference to knowledge held in other parcels (Lewandowsky, Kalish, & Ngang, 2002; Lewandowsky & Kirsner, 2000; Yang & Lewandowsky, 2003). In consequence, people may provide contradictory answers to a normatively identical problem, depending on which knowledge parcel they use to guide their answer. KP has been shown to arise with experts in domain-relevant tasks (Lewandowsky & Kirsner, 2000), with nonexperts in function learning (Kalish, Lewandowsky, & Kruschke, 2004; Lewandowsky et al., 2002), and with nonexperts in deterministic categorization tasks involving numeric (Yang & Lewandowsky, 2003) as well as various perceptual stimuli (Lewandowsky et al., 2006; Yang & Lewandowsky, 2004).

To illustrate, consider the study by Yang and Lewandowsky (2003), in which people learned to classify stimuli into one of two categories that were defined by two partial boundaries (i.e., boundaries that did not extend through the entire category space because people were only trained on a subset of items). The boundaries were at right angles to each other, and each bisected a unique segment in the two-dimensional space. During training, stimuli were accompanied by one of two "context" labels that were consistently mapped to the partial boundaries without, however, predicting category membership directly. Thus, context was a normatively irrelevant categorization cue, although it did predict which boundary could be used to classify a stimulus. At transfer, people who partitioned their knowledge (about one third of all participants) were found to rely exclusively on the boundary identified by context, even when classifying stimuli in a distant part of the category space. Moreover, people's performance within each context closely resembled the transfer performance of people in two control conditions who had only learned one of the partial boundaries. Thus, responding to an old stimulus in a *new* context was not influenced by learning in the original context, suggesting that partitioning was complete and the various parcels were independent of each other. Yang and Lewandowsky (2004) additionally showed that an exemplar model (Attention Learning COVering Map [ALCOVE]; Kruschke, 1992) was unable to account for the behavior of participants who partitioned their knowledge. Their performance was instead captured by a mixture-of-experts model that contained several independent rule modules (Attention To Rules and Instances in a Unified Model [ATRIUM]; Erickson & Kruschke, 1998).

Overall, KP has been firmly established as an attribute of deterministic category learning. It occurs equally with numeric stimuli (Yang & Lewandowsky, 2003) and with perceptual stimuli irrespective of whether they are perceptually integral or separable and irrespective of whether they are amenable to formulation of a simple verbal rule (Lewandowsky et al., 2006). In all instances, the context cue that gated use of knowledge by itself did not predict the outcome, $P(A|Context) = P(A)$, and context also did not alter the predictiveness of the remaining relevant cues; $P(A|Context, Relevant\ Cues) = P(A|Relevant\ Cues)$. Context was therefore not only irrelevant on its own but also did not constitute a compound of a conventional set of configural cues (see Yang & Lewandowsky, 2003, for a detailed analysis).

Notwithstanding the evidence for KP in deterministic categorization, there are several reasons why one might not expect it to occur in MCPL or probabilistic categorization. First, partitioning involves reliance on a normatively irrelevant cue; however, irrelevant cues are typically ignored in MCPL (Edgell et al., 1996; Kruschke & Johansen, 1999). Second, because there is evidence that learning in probabilistic environments gradually attenuates (Busemeyer & Myung, 1988), early learning strategies are likely to perseverate throughout the task. Early learning is known to rely on simple heuristics (e.g., one-dimensional rules; Nosofsky, Palmeri, & McKinley, 1994) that are replaced by other, more extensive

representations only later in learning (Johansen & Palmeri, 2002). Given that KP relies on the discovery of a correlation between context and other relevant cues, its emergence with probabilistic cues may therefore be less likely. Finally, the emergence of KP in probabilistic categorization appears particularly doubtful in light of related findings in causal learning (Young et al., 2000). When people simultaneously learn a positive patterning task (i.e., in which a compound cue AB predicts a positive outcome, but the components A and B each predict a negative outcome) and a negative patterning task (i.e., C and D each predict a positive outcome, but the compound cue CD predicts a negative outcome), people tend to use an "opposite" rule—that is, they learn that the compounds and their respective components have contrasting outcomes (Shanks & Darby, 1998; Young et al., 2000). However, when an additional, irrelevant probabilistic cue is present, use of the opposite rule is disrupted, and participants' performance is more consistent with the use of exemplars (Young et al., 2000). The opposite rule can be considered a rough analogue of KP (i.e., there are two associations that are applied in a contrasting manner in two different situations, both of which involve the same cues overall); hence, introducing probabilistic reinforcement in a categorization task might be expected to discourage KP.

That said, other evidence is suggestive of the possible occurrence of KP in MCPL. This evidence relies on the fact that people are sometimes sensitive to normatively irrelevant information. For example, people will use a cue on the basis of the strength of its pairing with an outcome, irrespective of the absolute frequency with which the two outcomes occur. In consequence, people will reliably use a cue that is rendered normatively irrelevant by the differing base rates of two outcomes (Gluck & Bower, 1988; Kruschke, 1996). To our knowledge, all recorded instances of irrelevant cue use in MCPL to date have involved neglect or misuse of base-rate information (Estes, 1976; Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Gluck & Bower, 1988; Kruschke, 1996; Nosofsky, Kruschke, & McKinley, 1992; Shanks, 1990, 1991). Nonetheless, those instances raise the possibility that people may also use a nonpredictive cue to partition their knowledge.

Should KP occur in probabilistic categorization, this would constitute a theoretically important outcome for at least two reasons. First, it would be the first demonstration that people sometimes "irrationally" use an irrelevant cue during MCPL when base rates are equal. The utilization of an irrelevant cue to gate access to different knowledge would present a problem for attention shifting mechanisms used in exemplar models, such as GCM (McKinley & Nosofsky, 1995) and RASHNL (Kruschke & Johansen, 1999), as these mechanisms are designed to shift attention only to the most predictive dimensions. Second, and perhaps most important, the finding would suggest a need for theories that rely on heterogeneous representations and that acknowledge the simultaneous coexistence of several partial knowledge components. None of the extant theories are likely to meet those requirements.

We now present two experiments in which we explored whether KP can occur in MCPL. In all experiments, quasi-continuous cues predicted a discrete outcome that was probabilistically reinforced. Cues were arrayed along an ordinal scale and were accompanied by a context cue that by itself did not predict the outcome and also did not enter into any configurally predictive combinations.

BEHAVIORAL EXPERIMENTS

Experiment 1

The primary goal of Experiment 1 was to determine whether KP can occur in an MCPL task. In this experiment, we used one relevant quasi-continuous cue (instantiated as the degree of shading of a colored disk) and one irrelevant binary context cue (the color of the stimulus). People were trained to classify stimuli into one of two categories.

To facilitate KP, the probabilistic assignment of the target outcome to the relevant cue increased from .20 to .80 as the shading of the stimulus increased in one context, whereas in the other context, the target probabilities decreased from .80 to .20 as the shading increased further (see Table 1). The overall probability of the target was thus identical between contexts. The degree of shading, irrespective of color, was by itself entirely predictive of the target probabilities, albeit nonlinearly.

Following training, participants were presented with stimuli comprised of all combinations of shading and color. If people partition their knowledge, they should associate each context with a partial probability function and use each partial function to respond to all stimuli presented in that context. It follows that KP would be manifest if people generalized along the relevant partial

Table 1
Stimulus Structure Used in All Experiments

Variable										
					Experiment 1					
Stimulus	1	2	3	4	5	6	7	8	9	10
Shading[a] (%)	0	10	20	30	45	55	70	80	90	100
Context		1	1	1	1	2	2	2	2	
P(A)	T[c]	.2	.4	.6	.8	.8	.6	.4	.2	T
					Experiment 2					
Stimulus	1	2	3	4	5	6	7	8	9	10
Numerosity[b]	1	2	3	4	5	6	7	8	9	10
Context		1	1	1	1	2	2	2	2	
P(A)	T	.2	.4	.6	.8	.8	.6	.4	.2	T

[a] Shading refers to the color of the circles used as stimuli, where 0% would be a completely unshaded circle (border only), and 100% would be a completely filled circle. [b] Numerosity refers to the number of circles presented. [c] Stimulus dimension values that are only shown at transfer are marked as "T."

function when presented with stimuli of the same color outside the trained range. Conversely, if people do not partition their knowledge, their responses should be identical across contexts. To maximize diagnosticity and facilitate modeling, we introduced two items outside the trained range of the continuous dimension at transfer (see Table 1).

Method

Participants

Twenty undergraduate psychology students from the University of Western Australia (Crawley, Australia) received partial course credit for participation.

Stimuli and Apparatus

Participants were trained individually on a Windows PC that presented stimuli and recorded responses using a MATLAB program written using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1991).

Each stimulus consisted of a single circle that varied in shading and color. Shadings were created with Microsoft Word's fill effects, with gradients of 10%, 20%, 30%, 45%, 55%, 70%, 80%, and 90% (see Table 1). Stimuli were presented on a white background in either red or green. For each participant, one color ("context" in Table 1) was randomly assigned to the increasing partial probability function, with the other assigned to the decreasing segment. During transfer, all levels of shading were shown in both colors, eight of which were new because they involved a novel combination of context and shading. Additionally, the two extreme gradients (no shading and full shading) were withheld during training and only shown at transfer, yielding 10 unique transfer items.

Design and Procedure

Training consisted of eight blocks of trials, each involving five presentations of the eight training items in a different random order, for a total of 320 trials. On each training trial, a randomly chosen stimulus was presented, and participants had to assign one of two possible outcomes by pressing the "F" or "J" key (assignment of keys to outcome was randomized across participants). Each stimulus was presented as an "alien cell" that was to be diagnosed as a "Fes-tins" or a "Jun-gins," with those labels being randomly assigned to the abstract A/B outcomes. Each response was followed by feedback ("CORRECT" or "WRONG"), generated randomly according to the probabilities in Table 1, presented underneath the stimulus for at least 1 s, after which participants pressed the spacebar to advance to the next trial. In addition, the percentage correct was shown at the end of each training block.

Transfer trials were identical to training trials except that feedback was withheld and a 750-ms blank interval followed each response.

Results

Training Performance

To identify people who were unable to learn the task, we computed the root-mean-square deviation (RMSD) between each participant's proportion of target responses during the final two training blocks and chance performance across all items. One participant whose RMSD was below .15 (two standard deviations below the mean; this cutoff was also used in Experiment 2) was excluded from the analyses.

Training performance was assessed by computing a probability matching score (PM score) that summarizes performance relative to the actual training probabilities for each participant:

$$PM_{ij} = (P(A|j) - R_i(A|j)) \times SI_j \quad (1)$$

where $P(A|j)$ is the training probability shown in Table 1 for item j, $R_i(A|j)$ is participant i's proportion of Category A responses for item j, and SI_j is the *signed indicator* of item j (i.e., $SI_j = +1$ if $P(A|j) > .5$, and $SI_j = -1$ if $P(A|j) < .5$; see, e.g., Friedman & Massaro, 1998). Scores were averaged across all items to compute each participant's PM score. A PM score of zero indicates perfect probability matching, whereas a negative score indicates overshooting of the training probabilities (with a PM score = –.25 indicating perfect maximizing), and a positive score indicates undershooting of the training probabilities (with a PM score = .25 indicating chance performance, and a PM score > .25 indicating a reversal of the training probabilities).

Table 2 shows that the average PM score tended toward zero (the separate groups displayed in Table 2 differentiate between participants on the basis of transfer performance and are explained below). Figure 1 provides visual confirmation that people learned to match the probabilities of the target outcome.

Table 2
Average Performance Measures for the Final Two Training Blocks From All Experiments

	RMSD		PM score		Consistency		
Experiment type	*M*	*SD*	*M*	*SD*	*M*	*SD*	*n*
Experiment 1	0.36	0.08	0.00	0.12	0.49	0.44	19
CI group	0.38	0.07	−0.05	0.07	0.42	0.51	7
KP group	0.40	0.06	−0.05	0.13	0.58	0.37	6
Experiment 2	0.31	0.10	0.07	0.12	0.51	0.42	36
CI group	0.36	0.09	0.01	0.12	0.56	0.33	11
KP group	0.29	0.09	0.04	0.13	0.38	0.48	10

Note. RMSD = root-mean-square deviation; PM = probability matching; CI = context-insensitive performance; KP = knowledge partitioning.

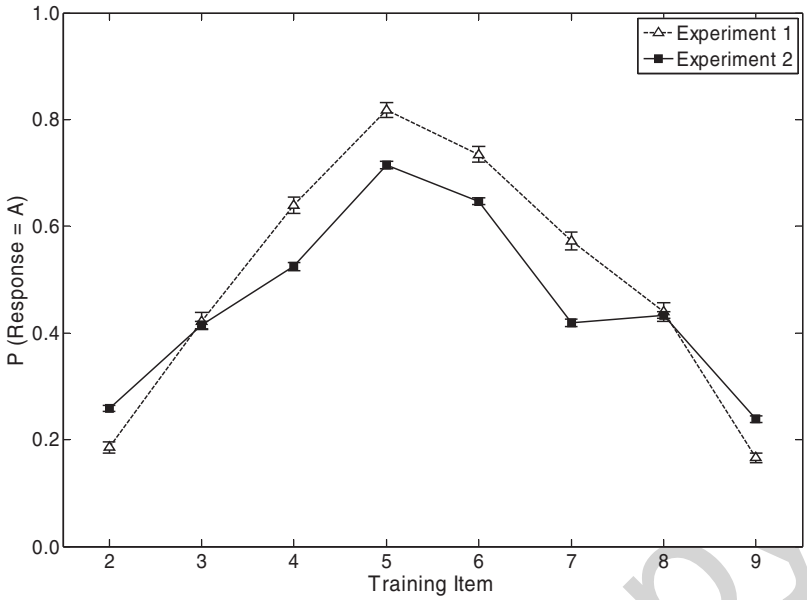

Figure 1. Mean proportion of Category A responses (with standard error bars) averaged across participants in the final two training blocks for Experiments 1 and 2.

Transfer Performance

Aggregate analysis. As a first step, the success of the context manipulation was examined by submitting the aggregate transfer data (converted to difference scores by subtracting responses in one context from responses to the same shading value in the other context) to a multilevel regression analysis with shading as the single predictor. A multilevel regression permits analysis of all data without confounding between- and within-subjects variability (e.g., Farrell & Lewandowsky, 2004; Lewandowsky & Brown, 2005). This analysis identified shading as a significant predictor of the differences between contexts, $F(1, 12) = 12.28$, $p < .05$, indicating that people overall were sensitive to context and that context-sensitivity interacted with shading. However, consistent with previous KP studies (see, e.g., Yang & Lewandowsky, 2003, 2004), inspection of the transfer data revealed clear individual differences in the way people approached the task; hence, the remaining analyses focused on these individual differences.

Cluster analysis. Following relevant precedent (Yang & Lewandowsky, 2003, 2004), we conducted a *k*-means cluster analysis on the individual profiles of transfer responses to identify common strategies among subgroups of participants. This analysis takes a set of starting points for *k* centroids ($k = 3$ in this case) and iteratively computes a solution that maximizes the between-clusters squared Euclidean distance while minimizing within-cluster distances. The clusters identified by the analysis corresponded to context-insensitive performance (henceforth CI; $n = 7$), KP ($n = 6$), and chance or idiosyncratic responding (e.g., if red, say "A"; otherwise, say "B"; $n = 6$). Participants in the latter cluster were not considered further.[2]

The transfer performance of the CI and KP groups is shown in Figure 2. People in the CI group clearly ignored context and responded only on the basis of shading. By contrast, people in the KP group responded very differently depending on the color of the stimuli.

To examine whether training performance might identify which transfer strategy people were acquiring, we performed independent-samples *t* tests between the two groups using various indicators of training performance. Neither the mean RMSD during training, $t(11) = -0.73$, $p > .05$, nor PM scores, $t(11) = -0.17$, $p > .05$, were found to differ (see Table 2). As a final comparison, participants' *consistency* of responding to the old training items during transfer was assessed by correlating each participant's aggregate responses during the final two training blocks with their responses to the same items at transfer. Table 2 shows that the KP group had a somewhat higher rate of consistency than the CI group, but this difference was not significant, $t(11) = -0.57$, $p > .05$.

Transfer strategies. Statistical exploration of the different strategies between groups involved separate 2 (Context) × 10 (Shading) within-subjects analyses of variance for each group. For the CI group, the analysis only revealed a significant effect of Shading, $F(9, 54) = 7.08$, $MSE = 0.18$, $p < .01$, $\eta_p^2 = .54$. No other effects were significant, with the largest $F(1, 6) = 2.35$ for the main effect of context. The main effect of Shading reflects a higher proportion of A responses for Items 4–7 than for the other items and also the increased response proportion for Item 1 com-

[2] Because a fair number of participants failed to be assigned to either of the groups of interest (CI and KP) by the *k*-means analysis, thus eliminating data from consideration, we also conducted an analysis that forced all participants into either the CI or KP groups (see, e.g., Yang & Lewandowsky, 2003, 2004). This analysis, which used the data from all participants, replicated 90% of all statistical effects reported for the two experiments on the basis of *k*-means clustering. It follows that none of our conclusions are altered if all participants are included in all analyses. Because modeling is facilitated by focusing on participants who most clearly differ from chance (and thus are most likely to differentiate between models), we only report the data based on the *k*-means analysis.

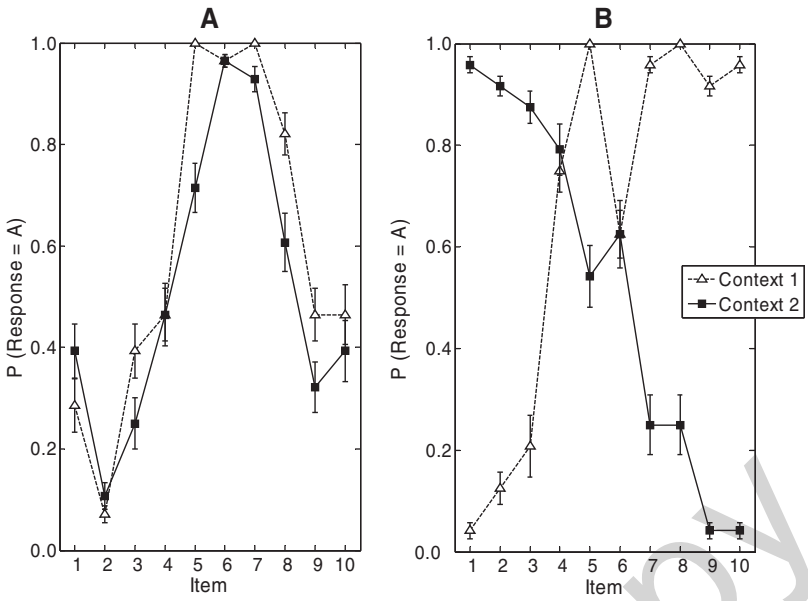

Figure 2. Mean proportion of Category A responses (with standard error bars) averaged across participants in the transfer phase of Experiment 1 for the (A) context-insensitive performance ($n = 7$) and (B) knowledge partitioning ($n = 6$) groups.

pared with Item 2 (accompanied by a lesser upturn for Item 10 compared with Item 9).

By contrast, the KP group modulated the use of shading on the basis of context. The significant Context × Shading interaction, $F(9, 45) = 29.55$, $MSE = 0.05$, $p < .01$, $\eta_p^2 = .86$, was accompanied by a significant main effect of Shading, $F(9, 45) = 2.58$, $MSE = 0.05$, $p < .05$, $\eta_p^2 = .34$, and a nonsignificant effect of Context, $F(1, 5) = 2.89$, $p > .05$. The main effect of Shading was due to Items 4–7 being categorized as A more frequently than the other items, whereas the interaction captured the cross-over extrapolation (see Figure 2).

To confirm the difference in extrapolation between the CI and KP groups, we tested for a linear trend in the difference scores between contexts. This analysis revealed a linear trend for the KP group, $F(9, 50) = 23.80$, $MSE = 0.13$, $p < .01$, $\eta_p^2 = .81$, but not for the CI group, $F(9, 60) = 0.21$, $p > .05$.

Discussion

In Experiment 1, we sought to determine whether people could be encouraged to partition their knowledge in an MCPL task by providing a normatively irrelevant cue that identified two subsets of cues whose relationships to the outcomes differed. Although a subgroup of participants ignored context and only utilized the relevant shading dimension, as expected from prior MCPL work on cue utilization (see Kruschke & Johansen, 1999, for a review), clear evidence for KP was present in another subgroup of participants who extrapolated quite differently in each context. The proportion of people who partitioned their knowledge (roughly one third of participants who exceeded chance during training) was commensurate with previous results (Lewandowsky et al., 2006; Yang & Lewandowsky, 2004). The current studies contrast with prior work by revealing a clear distinction between the CI and KP group's performance on the novel extrapolation items. The CI group did not extrapolate responses to the novel items from the relationship among trained items; instead, the novel items were categorized within the range of probabilities on which the CI group had been trained. The KP group, by contrast, exhibited systematic extrapolation outside of the training region.

The observed differences in strategy were not related to training performance and only became apparent at transfer, which again replicates work in deterministic categorization (Yang & Lewandowsky, 2003, 2004). Experiment 1 thus provided an existence-proof for KP in MCPL, thereby extending the precedents set in deterministic categorization (e.g., Yang & Lewandowsky, 2003, 2004) and function learning (Kalish et al., 2004; Lewandowsky et al., 2002).

Additionally, Experiment 1 demonstrated that KP was characterized by extrapolation outside the trained probability region. To ensure that the extrapolation observed in the KP group was due to learning and was not due to people's a priori "theories" regarding the likely relationship between shading in each context and the probabilistic feedback, we asked another 45 participants to assign to each transfer stimulus the likely probability of belonging to a single category in the absence of any training. The overwhelming response of those 45 participants was to ignore context altogether and to assign increasing probabilities to increasing shading values. A linear regression on all 45 participants' responses revealed a positive weight for shading ($b = 4.35$, $p < .01$) but not for context ($b = -0.17$, $p > .05$). Only two respondents utilized both shading and context, and both created a response rule that linked color and shading in a linear way. Interestingly, a small number of participants ($n = 5$) responded in a manner visually consistent with the responses of the CI group (i.e., higher probabilities for the middle shading values and smaller probabilities for the extreme values; a

quadratic regression for these individuals revealed significant regression weights for shading, $b = 43.17$, $p < .01$, and the squared shading term, $b = -3.70$, $p < .01$). The fact that no participants partitioned the stimulus space without prior training implies that partitioning arises as a part of the learning process and, by implication, suggests further exploration of our results through computational modeling. Before turning to the comparison of computational models, we first seek to replicate the findings from Experiment 1 and to enhance the KP effect by making the ordinal relationship between the continuous dimension and the training probabilities more salient.

Experiment 2

In this experiment, we asked whether the nature of the relevant cue affects performance. Any manipulation that makes the ordinal relationship between cues more salient and more readily discriminable might be expected to increase the prevalence of extrapolation and, by implication, the likelihood that people partition their knowledge. In Experiment 2, we therefore varied the numerosity of stimulus elements. Numerosity, given unlimited inspection time as in the current experiments, can be expected to result in unambiguous identification of the ordinal value of the stimulus. Hence, demonstration of KP with numerosity will provide additional evidence of the generality of the results.

Another exploratory purpose of Experiment 2 was to seek potential predictors for people's choice of strategy. At this point, we cannot anticipate whether an individual will exhibit KP or CI on the basis of training performance. All summaries of training performance were similar between both groups in the first experiment. An alternative possibility is that differences in strategies are precipitated by stable individual traits. In deterministic category learning, it has been shown that working memory capacity is related to KP, with partitioning being associated with lower working memory span (Yang, Lewandowsky, & Jheng, 2006). Given the undisputed connection between working memory capacity and general intelligence (Oberauer, Schulze, Wilhelm, & Süβ, 2005), it may likewise be the case that other variables related to general intelligence predict whether a person partitions his or her knowledge in MCPL. We explored this possibility by administering a digit symbol rotation task (Gignac & Vernon, 2003) that is known to correlate with general intelligence.

Method

Participants

Forty undergraduate psychology students at the University of Western Australia received partial course credit or remuneration (about $10 per hour) for participation.

Stimuli and Procedure

The stimuli and apparatus were identical to Experiment 1, with the exception that shading of a single circle was replaced by the simultaneous display of a varying number of fully shaded circles (see Table 1). In all other respects, the procedure was the same as in Experiment 1 with one exception: Following the MCPL training and transfer tasks, participants completed a digit symbol rotation task.

The digit symbol rotation task is a pencil-and-paper test in which people must mentally rotate and then reproduce arbitrary symbols that are mapped onto the Digits 1–9. The test sheet contains a key at the top with the 9 digits and their corresponding symbols. Underneath the key are five rows of target digits, each printed above an empty response box. For each target digit, participants identify the associated symbol in the key and reproduce it, rotated by 180°, in the response box. Participants were not permitted to rotate the test paper, thus requiring mental rotation of the symbols before reproduction. Participants were given 1.5 min to complete as many symbols as possible. The number correct on the digit symbol rotation task has previously been shown to load highly with a general intelligence factor (.63; Gignac & Vernon, 2003).

Results

Training Performance

Four participants were excluded from analysis on the basis of the RMSD cutoff. The remaining participants again learned to match the target probabilities (see Figure 1). Consistency scores across all participants were comparable with Experiment 1 (see Table 2).

Transfer Performance

Aggregate analysis. We again examined the success of the context manipulation by a multilevel regression using the difference scores between contexts as the dependent measure and stimulus numerosity as the predictor. Across all participants, numerosity again emerged as a significant predictor, $F(1, 20) = 13.62$, $p < .05$, indicating that the context manipulation was successful overall. As in the previous experiment, large individual differences were again present that were followed up by a cluster analysis.

Cluster analysis. Transfer performance of the groups identified by the *k*-means cluster analysis is shown in Figure 3 (CI, $n = 11$; KP, $n = 10$; chance, not considered further, $n = 15$). Independent samples *t* tests comparing the two groups of interest (KP vs. CI) on mean RMSD during training, $t(19) = 1.77$, $p > .05$, PM scores, $t(19) = 0.42$, $p > .05$, and consistency scores, $t(19) = 0.43$, $p > .05$, found no significant differences (see Table 2).

As in Experiment 1, the CI and KP groups were analyzed with two 2 (Context) × 10 (Numerosity) within-subjects analyses of variance. The significant Context × Numerosity interaction, $F(9, 81) = 6.69$, $MSE = 0.11$, $p < .01$, $\eta_p^2 = .43$, in the KP group, and the significant main effect of Numerosity, $F(9, 90) = 8.20$, $MSE = 0.22$, $p < .01$, $\eta_p^2 = .45$, in the CI group confirmed that Experiment 2 successfully replicated the preceding study. As in Experiment 1, the CI group exhibited an upturn in response proportions for the extrapolation Items 1 and 10.

In the KP group, items in Context 1 received on average a higher proportion of A responses than items in Context 2, resulting in a significant main effect of Context, $F(1, 9) = 5.58$, $MSE = 0.11$, $p < .05$, $\eta_p^2 = .38$. The largest of the remaining nonsignificant effects was the main effect of Numerosity in the KP group, $F(9, 81) = 1.62$, $p > .05$. As in Experiment 1, a comparison of the difference scores between contexts revealed a linear trend for the KP group, $F(1, 90) = 53.69$, $MSE = 0.22$, $p < .01$, $\eta_p^2 = .37$, but not for the CI group, $F(1, 100) = 0.52$, $p > .05$.

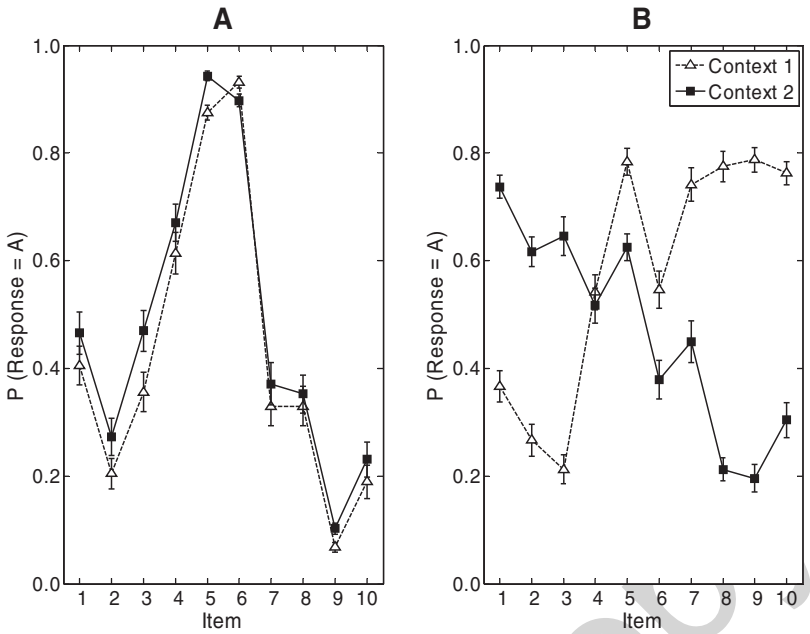

Figure 3. Mean proportion of Category A responses (with standard error bars) averaged across participants in the transfer phase of Experiment 2 for the (A) context-insensitive performance ($n = 11$) and (B) knowledge partitioning ($n = 10$) groups.

Digit Symbol Rotation

The rotation scores were first aggregated across participants, yielding a mean total number correct ($M = 25.16$, $SD = 9.33$) comparable with prior research (Gignac & Vernon, 2003). One participant performed more than two standard deviations above the mean; this participant was excluded from the correlational analysis. Point-biserial correlations were then computed between a participant's group membership (CI = 0, KP = 1) and that person's number of correct responses. Group membership was significantly and negatively correlated with the number correct ($r = -.56$, $p < .01$), indicating that the KP group ($M = 19.73$, $SD = 7.43$) had lower digit rotation scores than the CI group ($M = 30.00$, $SD = 8.38$).

Discussion

Using numerosity to represent the relevant cue did not affect the relative proportions of CI and KP participants. Experiment 2 thus largely confirmed the findings of the first study.

The discovery that KP was correlated with performance on a task that is known to be related to general intelligence adds support to the emerging finding that stable individual traits, such as intelligence or working memory capacity (Yang et al., 2006), are linked to the use of partitioning strategies. A corollary of this relationship is that KP should not be found in tasks that are exceedingly simple; this has indeed been shown to be the case in function learning (Lewandowsky et al., 2002) and categorization (Lewandowsky et al., 2006). We return to these ideas in the General Discussion section.

We now seek to pinpoint the underlying representations. If the CI and KP groups are using different strategies and are employing different underlying representations, then these differences should be identifiable through comparison of different computational models. Conversely, if the CI and KP groups can be adequately explained by a single common model, then the analysis should reveal the psychological mechanism responsible for their divergent transfer performance.

COMPUTATIONAL MODELING

We applied two exemplar models and a family of rule-based models to the data from both experiments. The exemplar models were the GCM (e.g., McKinley & Nosofsky, 1995; Nosofsky, 1986; Nosofsky & Johansen, 2000; Rouder & Ratcliff, 2004) and RASHNL, which was designed specifically to handle probabilistic categorization and MCPL (Kruschke & Johansen, 1999). The rule-based models were implemented within GRT (e.g., Ashby & Lee, 1991; Ashby & Maddox, 1993), which has successfully captured performance in probabilistic categorization (e.g., Rouder & Ratcliff, 2004).

We first optimized parameters with respect to the training data and obtained transfer predictions from each model. These predictions reveal the model's inherent properties based on the limited set of stimuli shown during training. We then additionally fit each model to the transfer data of the CI and KP groups separately. These fits ascertain whether a given model can handle the two transfer patterns by fine-tuning of parameters. To foreshadow our conclusions briefly, we found that the GCM provided the best fit for the CI group and that a variant of the GRT in which knowledge was explicitly partitioned provided the best account of the KP group.

GCM

The GCM postulates that people remember all previously seen items and base their judgments on similarity comparisons between the test item and all stored exemplars. This similarity comparison takes the form of

$$s_{ij} = \exp(-c \cdot d_{ij}), \tag{2}$$

where d_{ij} is the distance, in psychological space, between items i and j, and the specificity parameter, c, determines the steepness of the exponential similarity gradient (Nosofsky, 1986). The GCM captures a wide range of data by differentially allocating attention to different stimulus dimensions in the distance equation:

$$d_{ij} = \left(\sum_k w_k \cdot \left| x_{ik} - x_{jk} \right|^r \right)^{\frac{P}{r}}, \tag{3}$$

where x_{ik} is the value of dimension k for the test item i, x_{jk} is the value of dimension k for the stored exemplar j, w_k is the attention weighting for dimension k, r indicates the distance metric ($r = 1$ is a city-block distance that is primarily used for stimuli with separable dimension, and $r = 2$ is a Euclidian distance metric used for integral stimuli; Nosofsky, 1986), and P determines the form of the generalization gradient ($P = 1$, exponential; or $P = 2$, Gaussian; Shepard, 1987). For all simulations, r and P were set equal to 1. Additionally, only the attention weight to context, w_k, was estimated; attention to the relevant (shading or numerosity) dimension was $1 - w_k$.

Similarities are converted to response probabilities by applying Luce's choice rule (Luce, 1963):

$$P(A \mid i) = \frac{\left(\sum_{j \in A} M_{jA} \cdot s_{ij} \right)^{\gamma}}{\left(\sum_{j \in A} M_{jA} \cdot s_{ij} \right)^{\gamma} + \left(\sum_{j \in B} M_{jB} \cdot s_{ij} \right)^{\gamma}}, \tag{4}$$

where M_{jA} is the relative frequency with which a stored exemplar j is experienced together with the target outcome A, and M_{jB} is its frequency of occurrence with the other outcome (Cohen, Nosofsky, & Zaki, 2001). The relative frequency, M, was set equal to the frequency of presentation, such that M_{jA} corresponded to the array {8, 16, 24, 32} for Items 2–5, and the reverse for Items 6–9 (with the complement forming the arrays for M_{jB}). The response scaling parameter, γ, allows responding to vary between probability matching when $\gamma \cong 1$ and maximizing when $\gamma >> 1$ (Ashby & Maddox, 1993; Nosofsky & Johansen, 2000).

RASHNL

RASHNL (Kruschke & Johansen, 1999) was developed specifically for probabilistic categorization and was derived from the ALCOVE architecture (Kruschke, 1992), which instantiated the GCM within a connectionist network. RASHNL adds a rapid attention-shift mechanism and attenuation of learning rates onto the ALCOVE architecture. The rapid attention-shift captures the idea that when people have learned something about a task and then make an error with a new stimulus, they rapidly (but fleetingly) shift attention to the novel aspects of that stimulus. The attenuation of learning rates captures the fact that because error necessarily persists during probabilistic categorization, people must eventually discount it for learning to stabilize.

In RASHNL, each stimulus dimension has one real-valued input unit, and there is one exemplar node for each stimulus and one output node for each category. The weights between input and exemplar nodes are fixed and represent the location of the exemplar in a psychological space, the shape of which is determined by parameters representing the relative salience of each dimension. The weights between exemplar and category nodes are modified by an error-driven connectionist learning rule. Although the input-to-exemplar weights are fixed, each input dimension has an attention strength that is rapidly shifted during each trial.

Each exemplar node is activated to an extent determined by the combined effects of attention, salience, and the distance of the stimulus from that exemplar node. The attention given to a dimension i, α_i, is determined by a set of underlying gains, ϑ (the use of gains enables normalization of attention and computation of the derivative of attention with respect to error):

$$\alpha_i = \exp(\vartheta_i) / \left(\sum_{\substack{in \\ i}} \exp(\vartheta_i)^{\Gamma} \right)^{\frac{1}{\Gamma}} \tag{5}$$

The summation is over all dimensions, and Γ is an attention-normalization constant. When Γ is set to unity, any increase in attention to one dimension requires an equal decrease in attention to all other dimensions. When Γ is greater than one, any increase in attention results in a smaller decrease in attention to the other dimension. The reverse is true when Γ is smaller than one; any increase in attention to one dimension results in a greater decrease in attention to the other dimensions. The gains shift rapidly following feedback, so as to reduce error, as determined by a shift rate parameter Λ. Following the shift, any remaining error is used to drive both associative learning in the weights connecting exemplars to categories (see below) and to make long-term adjustments in dimensional attention. Thus, the model first shifts attention and then learns about the shifted stimulus and learns a little of the shift itself.

The activation function for an exemplar in RASHNL is identical to the GCM's distance function with P and r set to 1 (see Equation 3) but includes an additional parameter representing the salience of each input dimension. The influence of the salience of dimension i (ε_i) is identical to that of attention but is unchanging, as it reflects a static relationship between the dimension and the perceiver. Salience and attention determine activation of each exemplar unit j as follows:

$$\alpha_j^{ex} = \exp\left(-c \sum_{\substack{in \\ i}} \alpha_i \varepsilon_i \left| \psi_{ji} - a_i^{in} \right| \right), \tag{6}$$

where c is the overall width of the receptive field of that exemplar unit (also known as the specificity), α_i^{in} is the value of the stimulus on dimension i, and ψ_{ji} is the location of the exemplar j on dimension i. The dimensional salience is fixed throughout an experiment, whereas the attention strengths vary both within each trial as the gains are shifted, and over the course of the experiment as the gains are learned.

The model is governed by six basic parameters: a learning rate for the associative weights (λ_A), a learning rate for the attention

gains (λ_G), a shift parameter for the attention strengths (Λ), the attention normalization constant (Γ), a specificity parameter that determines the extent of generalization between exemplar representations (c), and the dimensional saliences (ε). In the present simulations, the salience of context was fixed ($\varepsilon_{ctxt} = 1.0$), but the relative salience of the (quasi-) continuous cues (shading or numerosity), ε_d, was estimated from the data. Details about the learning mechanisms can be found in Kruschke and Johansen (1999).

RASHNL additionally assumes that in response to the inevitable persistence of error in MCPL, people slowly come to discount their mistakes and cease to learn—that is, they no longer shift attention strengths or modify associative weights and attention gains despite encountering error. This discounting is modeled by a decay of the learning and shift parameters as follows:

$$r(t) = 1/(1 + \rho \cdot t), \tag{7}$$

where r is the current readiness of the network to learn, t is the trial number, and ρ is the decay parameter. All learning and shift rates are multiplied by r on each trial.

As with other exemplar models, RASHNL's category nodes produce real outputs that must be mapped onto response probabilities. Following precedent (Kruschke, 1992), this mapping uses an exponentiated form of Luce's choice rule, which introduces an eighth parameter, ϕ, that scales activations into response probabilities. The activity of each category node k is given by

$$\alpha_k^{cat} = \sum_{\substack{ex \\ j}} w_{kj}^{cat}\, \alpha_j^{ex}, \tag{8}$$

where w_{kj}^{cat} are the learned weights connecting exemplar unit j to category node k. The mapping to the probability that Category A is chosen from k options is given by

$$P(A) = \exp(\phi\alpha_A^{cat}) / \sum_{\substack{cat \\ k}} \exp(\phi\alpha_k^{cat}). \tag{9}$$

GRT

GRT is a multivariate extension of signal-detection theory that relies on the idea of variable stimulus perception (Ashby & Townsend, 1986). The degree of perceptual overlap between stimuli affects whether a stimulus is identified correctly. For example, in one of our experiments, Item 2 might on some trials be perceived as Item 1, and on other trials, as Item 3. To model categorization, GRT assumes that participants set a boundary between the category centroids and respond according to which side of the boundary a stimulus is perceived to fall (Ashby, Ell, & Waldron, 2003; Ashby & Gott, 1988; Ashby & Maddox, 1993). Hence, the GRT explains probability matching by assuming that items close to the category boundary will be perceived as items of the opposing category more often than items far from the boundary.

GRT differs from GCM and RASHNL in at least two respects: First, GRT is not a single well-specified model but is best understood as a family of possible instantiations within a common architecture, each characterized by unique assumptions about the shape of the category boundary (e.g., Nosofsky, 1998). For example, in the present case, the two-dimensional (Context × Shading) category space could be divided by either one or by several linear boundaries, each described by a set of parameters that are estimated on the basis of people's categorization responses[3] (Nosofsky, 1998). In consequence, there is no consensus whether the boundaries are best estimated on the basis of performance on training items alone (e.g., McKinley & Nosofsky's, 1996, Experiment 1) or by considering novel transfer items as well (e.g., Maddox & Ashby, 1993; McKinley & Nosofsky's, 1996, Experiment 2). This issue is particularly relevant in the present case, in which there are more novel transfer items than training stimuli and in which the training stimuli cover only a small segment of the category space.

We implemented two variants of the GRT. The first model, called GRT-integrated, divided the two-dimensional space with two linear boundaries. The second model was an explicitly partitioned model, called GRT-partitioned, that replaced the two-dimensional space by two one-dimensional "slices," each associated with one of the contexts and each characterized by a single dimension representing shading. A separate boundary was estimated for each one-dimensional slice. Graphical representations of the boundaries are shown later, together with the transfer data, when the results are presented.

For each model, the locations of the boundaries were estimated as free parameters. Each boundary required one or two parameters, respectively, for the GRT-partitioned and GRT-integrated. The perceived shading (or numerosity in Experiment 2) of each item, i, was assumed to be normally distributed with a mean, μ_i, and variance, σ_i^2. Integrating over the density of this distribution in the appropriate region yields the predicted response probabilities (for details, see Ashby & Lee, 1991; Ashby & Maddox, 1993).

Parameter Estimation

For efficiency of presentation, the models were fit to the *combined* data of Experiments 1 and 2 by maximizing the log likelihood

$$\ln L = -n/2 \times \ln(SSD/n), \tag{10}$$

where n refers to the number of means being fitted, and *SSD* refers to the sum of squared deviations between those means and the predictions. As the data have been averaged first across item repetitions and then across subjects, the to-be-fitted means were assumed to be normally distributed; hence, we adopted a Gaussian probability model. Parameters were optimized with SIMPLEX.

All model comparisons were based on Akaike's information criterion (*AIC*; Akaike, 1974; see also Burnham & Anderson, 2002). The *AIC* corrects the log likelihood for the degrees of

[3] The GRT can also be instantiated with two quadratic boundaries dividing the two-dimensional (Context × Shading) category space. However, each of these boundaries requires five free parameters, and fitting the model to the training data alone resulted in boundaries that were essentially linear in form. Fitting the model to the transfer data implies that all stimuli, including those not seen at training, jointly determine placement of the boundary. We consider this to be psychologically questionable because it implies that people rely on either precognition or the extremely small, but arguably nonzero, probability that a quadratic boundary placed to accommodate the training ensemble will settle on exactly those parameter values that yield KP at transfer; hence, we restrict our presentation of the GRT to the linear versions.

Table 3
Model Fits to Combined Training Data From Experiments 1 and 2 and Fits to Combined Transfer Data Using Parameters Estimated From the Fit to the Training Data

	AIC_c (w_iAIC)					RMSD				
		Smooth transfer		Raw transfer			Smooth transfer		Raw transfer	
Model	Training data	CI	KP	CI	KP	Training data	CI	KP	CI	KP
GCM	−38.72 (0.00)	**−93.45 (1.00)**	−38.59 (0.00)	**−80.68 (1.00)**	−35.09 (0.00)	0.04	0.08	0.32	0.11	0.34
RASHNL	−164.97 (0.00)	−57.58 (0.00)	−21.52 (0.00)	−47.51 (0.00)	−18.10 (0.00)	0.10	0.11	0.28	0.15	0.31
GRT-integrated	**−181.19 (1.00)**	−55.25 (0.00)	−13.60 (0.00)	−45.67 (0.00)	−10.53 (0.00)	0.04	0.12	0.34	0.15	0.37
GRT-partitioned	−39.95 (0.00)	−31.28 (0.00)	**−63.43 (1.00)**	−29.01 (0.00)	**−57.85 (1.00)**	0.04	0.38	0.17	0.40	0.20
Chance performance						0.19	0.21	0.31	0.25	0.29

Note. Minimum Akaike's information criterion corrected for finite samples (AIC_c) among competing models for each group and data set are in bold. w_iAIC = Akaike weights; RMSD = root-mean-square deviation; CI = context-insensitive performance; KP = knowledge partitioning; GCM = generalized context model; RASHNL = rapid attention shifts 'n learning; GRT = general recognition theory.

freedom of the model, as reflected in the number of free parameters that must be estimated. Formally, the *AIC* (corrected for small samples; e.g., Wagenmakers & Farrell, 2004) is given by

$$AIC_c = -2 \ln L + 2K + \frac{2K(K+1)}{(n-K-1)}, \quad (11)$$

where K is the number of free parameters involved in maximizing ln L. The *AIC* thus trades off goodness-of-fit (ln L) against lack of parsimony (K) with an added correction recommended for cases in which the ratio of data points to parameters (n/K) is less than 40.

Fit to Training Data

We first fit the models to the training data during the last two blocks (i.e., the average of the two curves shown in Figure 1). As RASHNL models trial-by-trial learning, thus making varied predictions depending on the particular sequence of training items, predictions from RASHNL were generated by averaging the performance of 25 simulated participants, each with a different random training order and different random probabilistic sampling of reinforcement. The GCM and GRT do not model learning, and hence their fits were based on a single prediction.

The stimulus inputs for each of the models were the ordinal stimulus values (i.e., 1, 2, . . ., 10) shown in Table 1. Preliminary modeling indicated that replacing those values with a multidimensional scaling solution (for GCM and RASHNL; on the basis of similarity judgments for a monochrome version of the shading dimension used in Experiment 1) or a confusion matrix (for GRT) did not substantially improve the model fits and did not affect the conclusions; hence, we only report fits based on the untransformed input values.

Table 3 summarizes the fits of the models using AIC_c and Akaike weights (w_iAIC). The latter facilitate model comparison because they can be interpreted as the conditional probability that a given model i is the best model of the set being compared (Wagenmakers & Farrell, 2004).[4] Table 3 additionally shows the RMSD associated with each fit, which is directly interpretable as the deviation between predictions and data. All models provide a reasonable approximation of the training data, although the w_iAIC favors the GRT-integrated over all other models. To maximize the diagnosticity of the modeling, we assessed goodness-of-fit not only with respect to the observed response probabilities but also with respect to a smoothed data set created by computing a moving average with a window size of three stimuli across all nonextrapolation items. For the smoothed data set, the model predictions were also smoothed prior to calculating the fit.

The associated transfer predictions of the models are shown in Figure 4, with the estimated parameter values in the caption and goodness-of-fit statistics shown in Table 3. Recall that those predictions were based on parameter values that were optimized with respect to the training data only. Not surprisingly, the GRT-partitioned predicted the context-specific extrapolation observed in the KP group. The remaining models all predicted context-insensitive transfer performance, although only the two exemplar models either captured (RASHNL) or approximated (GCM) the upturn observed for novel items (i.e., Items 1 and 10).

The statistical summary in Table 3 confirms that the GCM best captured CI performance, whereas the GRT-partitioned was the only model to capture KP performance. All other models performed at chance or worse for the KP data.

We conclude that none of the models, except the one designed to partition knowledge, spontaneously produce KP performance. We next examine whether the models can be coaxed into exhibiting KP behavior by fitting them to the transfer data of both groups separately.

Fit to CI and KP Transfer Data

GCM

The predictions of the GCM when fit to the transfer data of both groups separately are shown in Figure 5 (parameter values of each model are shown in the relevant figure caption). Table 4 summarizes the fit statistics for all models.

[4] The log-likelihoods in the present case involve a probability density and can therefore exceed unity. In consequence, the AIC_cs that are negated in Equation 11 can be negative.

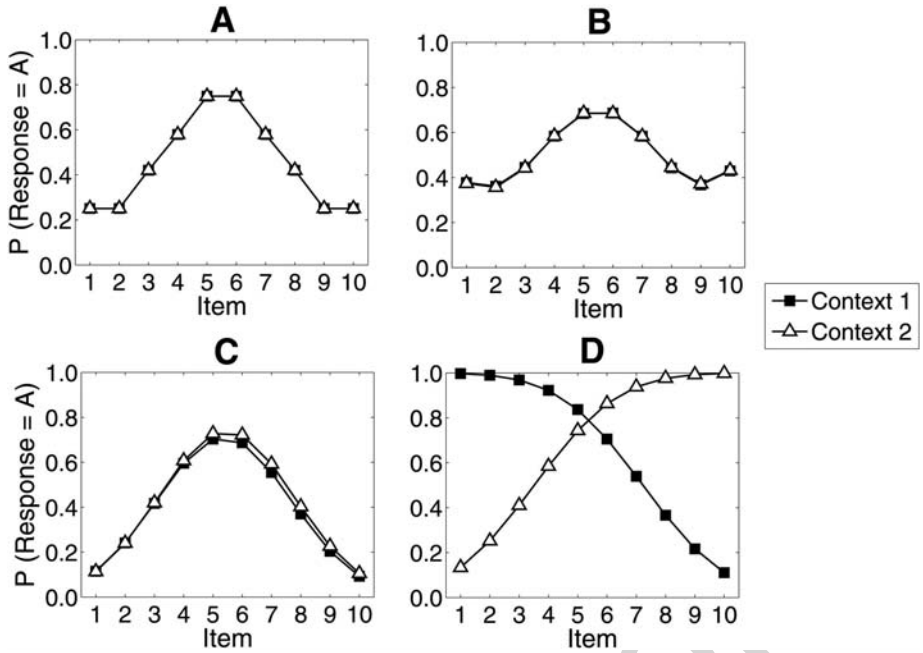

Figure 4. Model predictions for transfer stimuli, using parameter values estimated by fitting the combined training data of Experiments 1 and 2. Panel A shows predictions for the generalized context model ($c = 15.21$; $\gamma = 0.79$; $w_k = 0.00011$), Panel B for the rapid attention shifts 'n learning model ($c = 0.69$; $\Gamma = 0.96$; $\Lambda = 0.26$; $\lambda_A = 0.02$; $\lambda_G = 0.77$; $\phi = 0.96$; $\varepsilon_d = 1.00$; $\rho = 0.00000012$), Panel C for the integrated general recognition theory (GRT) with two linear decision bounds (Boundary 1: $-0.01x - 0.45y + 1.97$; Boundary 2: $0.01x + 0.05y - 0.36$; $\sigma^2_{ctxt} = 0.37$; $\sigma^2_d = 4.85$), and Panel D for the partitioned GRT with two linear boundaries ($\beta_1 = 3.36$; $\beta_2 = 7.45$; $\sigma^2_d = 1.88$). Note that the variables x and y refer to the values of context and shading, respectively, for all GRT fits. See text for explanation of model parameters.

It is clear from the table that GCM provides by far the best fit of all models to the CI transfer. Nonetheless, Panel A in Figure 5 shows that although GCM approximates the upturn seen in the data for novel items, it fails to capture its full extent. The GCM cannot generate a large upturn because its generalization to new items is dominated by the nearest trained neighbor, thus preventing it from deviating much from the nearest trained probability.[5]

Panel B in Figure 5 shows that when fit to the KP transfer data, GCM can find parameter values that permit it to exhibit a form of partitioning. However, this apparent success is marred by several problems. First, because all GCM can do is to focus attention exclusively on context (with $w_k \cong 1$), it again predicts nearest neighbor generalization (albeit in a context-specific manner) rather than the extrapolation observed in the data.[6] Second, the fact that KP performance is modeled by removing (virtually) all attention from the one dimension that is diagnostic during training, and by shifting it onto the one that is irrelevant, reduces the plausibility of the model's account. In support, Panels C and D in the figure show the best-fitting estimates of the three parameters across 500 different replications with different random starting values (but convergence to a common maximum likelihood). Panel C shows that for the CI group, there is uniform convergence on a single best estimate for c and γ with some moderate variation in the final estimate for the attention devoted to context. Panel D, by contrast,

[5] Allowing the exponent, P, to fall below one in Equation 3 allows the GCM to exhibit the upturn for the extrapolation items. To date, psychological plausibility has only been ascertained for values of P of 1 or 2 (Shepard, 1987) that map, respectively, to an exponential or a Gaussian generalization gradient. However, as P approaches zero, the continuous distance between two stimuli is transformed into a quasi-binary step function, with old items eliciting a nonzero similarity, whereas all new items are considered to be equally (and virtually completely) dissimilar. This step function turns out to be related to the binary multiplicative similarity rule of the original context model (Medin & Schaffer, 1978). Substituting the multiplicative similarity function into the GCM and assuming only two levels of similarity, either matched or mismatched (with a high penalty for a mismatching dimension, mismatch parameter, $s < .01$), did allow the GCM to exhibit the upturn for the novel items but at the cost of a reduced fit to the trained items for both the observed response probabilities (RMSD = 0.14).

[6] The GCM's failure to extrapolate was not caused by "forcing" the GCM to attend to an irrelevant dimension. An experiment (not reported in detail here) using a single, linearly increasing monochrome shading dimension showed that participants ($n = 20$) extrapolated their responses to new transfer items outside of the training region just like the KP groups did in Experiments 1 and 2. When applied to the data from that study, the GCM was again unable to produce extrapolation responses outside of the region on which it had been trained. Hence, the fundamental limitation of GCM in the present design is an inability to extrapolate, not an inability to show forms of KP based on context.

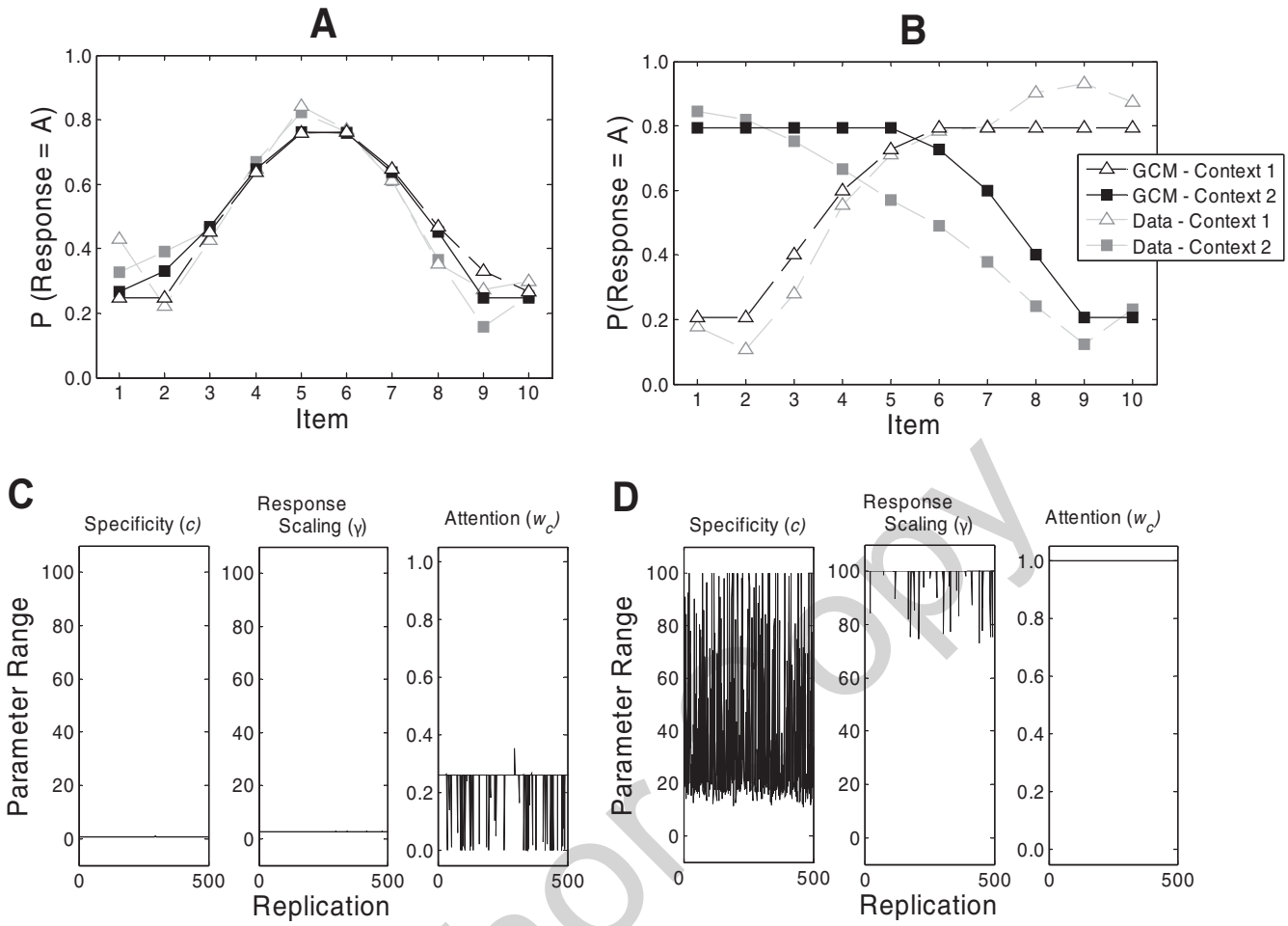

Figure 5. Generalized context model (GCM) fits to the combined smooth transfer data from Experiments 1 and 2 for (A) the context-insensitive performance (CI) group (c = 0.66; γ = 3.00; w_k = 0.25) and (B) the knowledge partitioning (KP) group (c = 33.60; γ = 100.00; w_k = 0.9996). Panels C and D show the parameters estimates across 500 independent parameter-estimation runs with different starting values for the CI and KP groups, respectively.

shows that for the KP group, attention is consistently (and exclusively) shifted onto context. This implausible removal of attention from the sole diagnostic dimension is, moreover, accompanied by large and highly variable estimates for c and γ. The large values of γ imply that all test items elicit nearly identical cumulative similarities (based only on items within the same context as per the large c) that are then boosted to match the appropriate probabilities by the large response scaling parameter.

Table 4
Model Fits for Combined Transfer Data From Experiments 1 and 2

	AIC_c (w_iAIC)				RMSD			
	Smooth transfer		Raw transfer		Smooth transfer		Raw transfer	
Model	CI	KP	CI	KP	CI	KP	CI	KP
GCM	**−101.18 (1.00)**	−78.32 (0.01)	**−81.96 (1.00)**	−69.11 (0.01)	0.07	0.12	0.11	0.15
RASHNL	−70.39(0.00)	−66.45 (0.00)	−56.89 (0.00)	−51.86 (0.00)	0.08	0.09	0.12	0.13
GRT-integrated	−66.99 (0.00)		−51.61 (0.00)		0.09		0.13	
GRT-partitioned		**−101.11 (0.99)**		**−78.37 (0.99)**		0.07		0.12

Note. Minimum Akaike's information criterion corrected for finite samples (AIC_c) among competing models for each group and data set are in bold. w_iAIC = Akaike weights; RMSD = root-mean-square deviation; CI = context-insensitive performance; KP = knowledge partitioning; GCM = generalized context model; RASHNL = rapid attention shifts 'n learning; GRT = general recognition theory.

RASHNL

The predictions of RASHNL are shown in Figure 6. Although the model describes the data well overall, its fit statistics are penalized by the large number of free parameters. In addition, like the GCM, the model shows little evidence of context-specific extrapolation for the KP group; instead, RASHNL's generalization gradients are quite flat and resemble those of the GCM. The fits to the CI group resemble the predictions obtained earlier when fitting the training data. RASHNL again captured the upturn for the novel items.

It is informative to analyze the model's behavior further. Recall that RASHNL's predictions were based on the average across 25 simulated subjects, each involving a different randomization of training stimuli. Panels C and D in the figure show the predictions for each of those simulated subjects for the CI and KP group, respectively. It is immediately clear that the CI group is simulated very consistently, as each of the individual graphs in Panel C resembles the aggregate predictions in Panel A. The picture is very different for the KP group: Although the aggregate predictions in Panel B mirror the data, this is rarely the case at the level of individual simulated participants (Panel D; each training sequence here was identical to the corresponding sequence in Panel C). With the set of parameters necessary to model KP, RASHNL becomes unduly sensitive to trivial random differences between training sequences.

A possible explanation for this sensitivity is that RASHNL's learning mechanisms are based on those in ALCOVE (Kruschke, 1992), which has been shown to be extremely sensitive to tiny changes in reinforcement during probabilistic category learning (Lewandowsky, 1995). Lewandowsky (1995) showed that a single change in a long sequence of probabilistic stimuli substantially altered the predictions of ALCOVE, and RASHNL

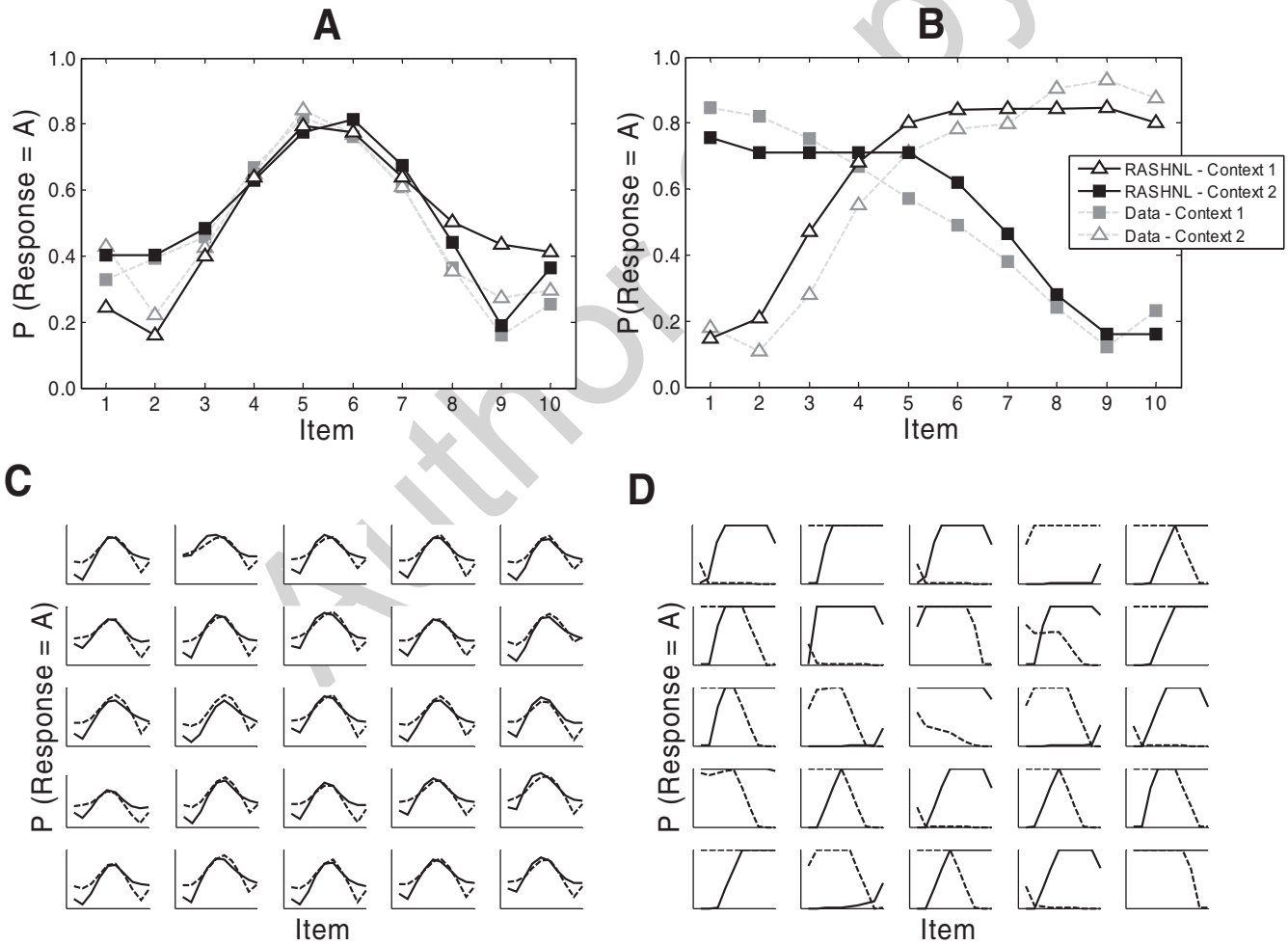

Figure 6. Rapid attention shifts 'n learning (RASHNL) fits to the combined smooth transfer data from Experiments 1 and 2 for (A) the context-insensitive performance (CI) group ($c = 2.80$; $p = .71$; $\Lambda = 0.00$; $\lambda_A = 0.00032$; $\lambda_G = 0.00$; $\phi = 100$; $\varepsilon_d = 0.94$; $\rho = 0.0000015$) and (B) the knowledge partitioning (KP) group ($c = 92.59$; $\Gamma = 6.56$; $\Lambda = 0.00$; $\lambda_A = 0.04$; $\lambda_G = 1.31$; $\varphi = 100$; $\varepsilon_d = 0.0023$; $\rho = 0.00$). The RASHNL predictions have been aggregated over 25 different random sequences of training items. Panels C and D show the 25 individual predictions for each training sequence for the CI and KP groups, respectively.

may be experiencing similar difficulty here (in particular because the final parameter estimates precluded attenuation of learning). In addition, as in the GCM, the response scaling parameter, ϕ, was very large, and the attention shift rate and learning rates were all approaching zero. These values suggest that little learning takes place and that the model's predictions were based on amplification (by the scaling parameter) of slight differences between exemplars.

GRT

For this final fit, the CI group was modeled by the GRT-integrated, whereas the KP group was modeled by the GRT-partitioned. The results are shown in Figure 7, together with each model's underlying representation of the category space and the boundary placements (Panels C and D). It is clear from the figure and Table 4 that the GRT-integrated handled the CI data well, although its account fell short of those provided by the exemplar models. This shortcoming largely reflected the failure of the GRT to predict an upturn for the novel items. Likewise, the GRT-partitioned provided a very good account of performance in the KP group.

Panels C and D show that the boundaries are placed in a manner consistent with the information available during training. In confirmation, these boundaries resemble those created during the earlier fit to the training data (compare parameter estimates in the captions of Figure 4 and Figure 7). It is informative that the account of the KP group required the complete encapsulation of partial boundaries within two separate "parcels" of knowledge, without any cross-talk or sharing

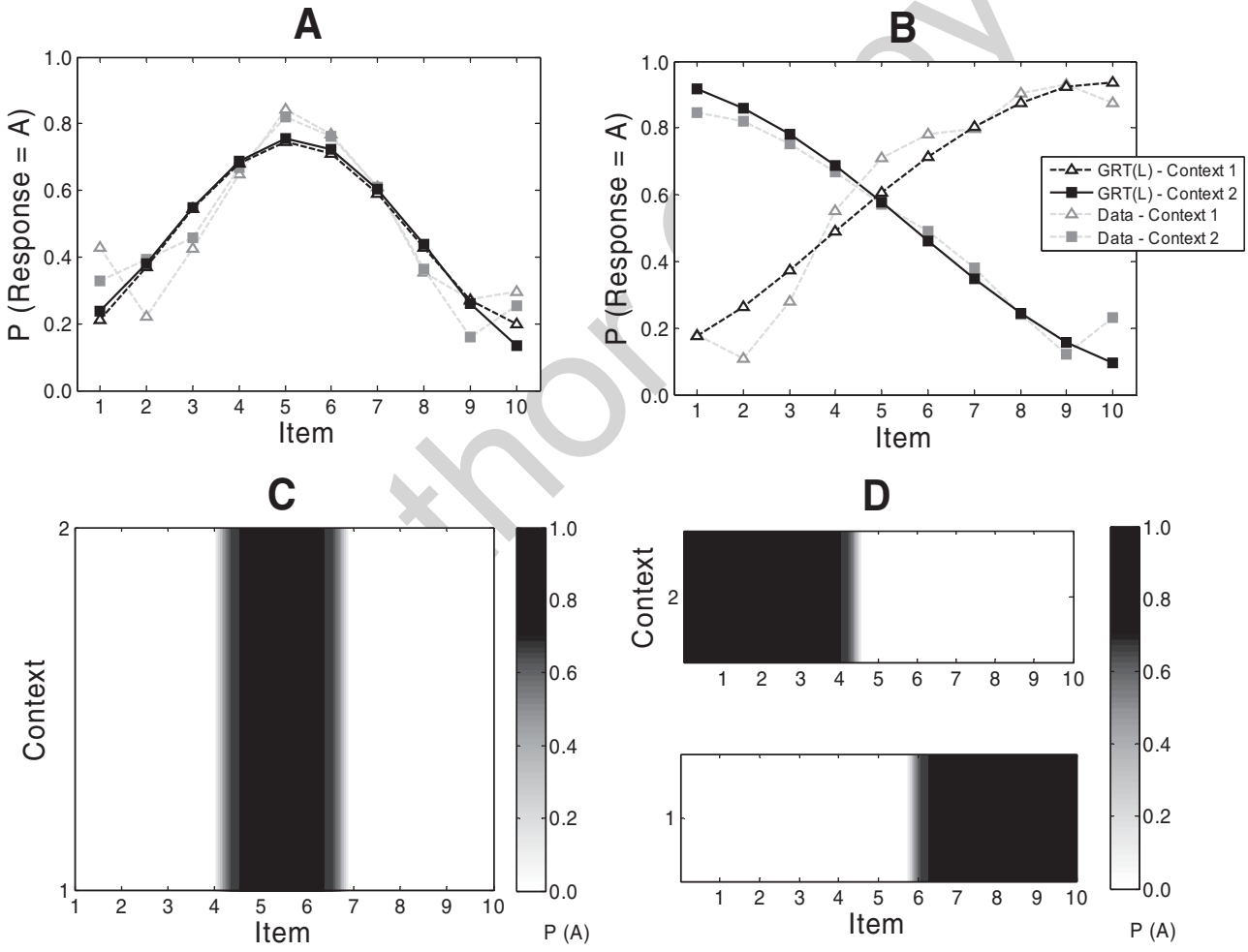

Figure 7. Panel A shows the integrated general recognition theory (GRT) with two linear (L) decision boundaries fit to the combined smoothed transfer data from Experiments 1 and 2 for the context-insensitive performance (CI) group (Boundary 1: $0.00x - 0.24y + 0.93$; Boundary 2: $-0.03x + 0.41y - 2.58$; $\sigma^2_{ctxt} = 0.09$; $\sigma^2_d = 6.04$). Note that x and y refer to the values of context and shading, respectively. Panel B shows the partitioned GRT with two L decision boundaries fit to the combined smoothed transfer data from Experiments 1 and 2 for the knowledge partitioning group ($\beta_1 = 3.64$; $\beta_2 = 5.85$; $\sigma^2_d = 2.07$). Panel C is the two-dimensional contour plot of the category space showing the predicted category boundaries for the CI groups. Panel D shows two partitioned unidimensional category spaces each with a single L boundary.

of information between parcels. This confirms the unique heterogeneity of representations that is required to accommodate KP performance. That said, notwithstanding the diagnostic value of the account by the GRT-partitioned, the model does not specify the processes by which people create those partitions.

Conclusions From Modeling

1. All models bar the GRT-partitioned responded in a context-insensitive manner when fit to the training data. None of those models "spontaneously" partitioned their knowledge when trained in the same manner as our participants.

2. When applied to the KP transfer data, both exemplar models captured some aspects of people's performance and produced a form of KP, though not of the type observed in the data. Neither the GCM nor RASHNL could predict the linear extrapolation outside the trained range that was observed for the KP group.[7] Moreover, further analysis of the models' behavior uncovered extraneous reasons (e.g., problematic parameter values and extreme sensitivity to stimulus sequences) to reject their account of KP. The fact that neither the GCM nor RASHNL provided a satisfactory account of the KP data suggests that it is the exemplar architecture, rather than model-specific assumptions, that prevent an account of partitioning.

By contrast, an instantiation of the GRT based on partitioned knowledge provided an excellent account of KP performance. The GRT-partitioned placed boundaries in a manner consistent with the information available during training, and it confirms the need to assume multiple independent knowledge components to accommodate KP performance, without however specifying how those partitions come about.

3. The GCM and RASHNL successfully accounted for CI performance, although the former was preferred on the basis of its parsimony. Both models to varying extents exhibited the upturn for novel items, which we therefore consider to be a signature of exemplar representations. The GRT-integrated also provided a plausible account of the CI group, although it failed to accommodate the upturn associated with novel items. Taken together, the behavior of those three models strongly implies that the CI group relied on exemplar representations, irrespective of how exactly those representations are instantiated in a model.

4. Perhaps most important, the modeling implies that homogeneous representations, as embodied in an all-exemplar model such as RASHNL or in an integrated variant of the rule-based GRT, are insufficient on their own to account for performance in MCPL. Instead, under identical circumstances, performance may rely either on exemplar representations or on multiple independent partial rules. Below we sketch a direction for future theorizing that conforms to these constraints.

GENERAL DISCUSSION

Concerns and Limitations

Before we discuss the implications of our findings, we take up some concerns and limitations. Two related concerns may be voiced about our results. First, one might question the use of a cluster analysis to identify subgroups of participants and, second, one might be concerned about the exclusion of people from the analysis on that basis. Our response is two-fold. First, we note that the aggregate multilevel regression identified a significant effect of the context manipulation in both experiments, thus allaying concerns that the distinct subgroups might have emerged post hoc, merely by dividing an otherwise noisy data set in convenient ways. Second, we remind the reader that an analysis based on a forced assignment of *all* participants to either the CI or KP group did not change any of the conclusions. We furthermore note that the analysis of distinct subgroups of participants has ample precedent (Juslin et al, 2003; Lee & Webb, 2005; Lewandowsky et al., 2006; Navarro, 2007; Navarro, Griffiths, Steyvers, & Lee, 2006; Nosofsky, Clark, & Shin, 1989; Nosofsky & Palmeri, 1998; Rouder & Ratcliff, 2004; Yang & Lewandowsky, 2003, 2004).

A second concern might be that we were not able to differentiate the two groups by consideration of their training performance. However, we were able to predict whether an individual will exhibit KP or will ignore context on the basis of performance on a task that is known to be related to general intelligence. The finding that people of higher ability are less likely to partition their knowledge is consonant with recent research in deterministic categorization (Yang et al., 2006). Yang et al. (2006) found that people with low working memory span were more likely to partition their knowledge in a deterministic categorization task than people with a high working memory span. Although these results are beginning to converge on the conclusion that partitioning is a response to excess complexity, the precise processes underlying this relationship presently remain unknown. This conclusion is also consistent with a recent analysis of concept complexity that distinguishes between when people can and cannot use rules or exemplar representations (Feldman, 2006); we return to this idea below.

Theoretical Implications

Status of KP

The finding that people can partition their knowledge in a probabilistic categorization task adds to a growing body of research pointing to the ubiquity of partitioning in concept learning (Kalish et al., 2004; Lewandowsky et al., 2002; Lewandowsky & Kirsner, 2000; Lewandowsky et al., 2006; Yang & Lewandowsky, 2003, 2004, 2005). The proportion of people who partitioned their knowledge in the present studies (29% overall, and nearly 50% of those who fell into one of the two groups of interest) was commensurate with the proportion observed in deterministic categorization (Lewandowsky et al., 2006; Yang & Lewandowsky, 2003, 2004).

Lewandowsky et al. (2006) showed that in deterministic categorization, partitioning transcended the putative distinction between separate cognitive systems advocated by Ashby and colleagues (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998),

[7] A related function-learning exemplar model (EXtrapolation-Association Model [EXAM]; DeLosh, Busemeyer, & McDaniel, 1997) can show extrapolation behavior by basing its responses on a local slope estimated from the stored exemplars. However, this model would require modification to produce a discrete response, and, furthermore, Kalish et al. (2004) have demonstrated that EXAM is unable to account for KP in function learning. Hence, we forego further discussion of this model.

one of which is dedicated to use of verbalizable rules. Unlike several previous dissociations depending on whether category rules were verbalizable (e.g., Ashby et al., 2003; Ashby & Waldron, 1999; Maddox & Ashby, 2004; Maddox, Bohil, & Ing, 2004), Lewandowsky et al. (2006) showed that this distinction did not eliminate KP. The present data show that KP similarly transcends a distinction—between deterministic and probabilistic categorization—that in many other instances has proved sufficiently powerful to provoke qualitative changes in generalization behavior (e.g., Mehta & Williams, 2002). Moreover, the existence of partitioning in MCPL is particularly intriguing because as noted at the outset, previous research had rendered its occurrence with probabilistic reinforcement rather unlikely.

The use of probabilistic categorization also permitted examination of KP when the task involved a single dimension of relevant cues and a very small number of training stimuli. In all previous examinations of partitioning, the partial rules involved two-dimensional boundaries (e.g., Yang & Lewandowsky, 2003), and the training ensemble was considerably larger. That KP persists in the current experiments with a single relevant cue and with few training instances attests to the fact that MCPL is in many ways quite different from deterministic categorization and function learning. We consider the implications associated with the small number of stimuli later.

Yang and Lewandowsky (2004) identified the importance of the mixture-of-experts approach in modeling KP but did not specify the forms of the specific modules that would be needed. In particular, Yang and Lewandowsky left open the possibility that a partitioned exemplar model (i.e., multiple independent modules of exemplars) might also be able to capture KP. The current data speak against this idea because an exemplar model, even if partitioned, could not show the linear extrapolation that was here associated with KP. Hence, the current results, which show that KP was only linked to rule-based representations and not to exemplar representations, provide a clear constraint for future mixture-of-experts modeling, a point considered further below.

Relationship to Probability Learning

The context cue in our experiments was normatively irrelevant by two strong criteria: It predicted nothing by itself, and its presence did not alter the predictiveness of another cue (which is the hallmark of configural compounds—i.e., when two individually irrelevant cues are jointly predictive, configural information is known to be used in MCPL; see Edgell, 1995). Notwithstanding its normative irrelevance, context was a primary determinant of performance for a significant proportion of participants. We argue that this reliance on an irrelevant cue differs considerably from previous demonstrations of "irrational" cue use in MCPL.

Gluck and Bower (1988; see also Cobos, López, Rando, Fernández, & Almaraz, 1993; Estes et al., 1989; Myers, Lohmeier, & Well, 1994; Nosofsky et al., 1992; Shanks, 1990) presented participants with a long sequence of classification trials on which two outcomes (call those R and F) had to be predicted on the basis of a number of cues. One of the cues (call that C) was present more often with outcome R than with outcome F; hence, $P(C|R) > P(C|F)$. However, because one outcome (R) was rare relative to the other more frequent one (F), the cue C was normatively irrelevant to deciding which outcome was present; hence, $P(F|C) = P(R|C)$. Nonetheless, on a final classification test, people preferred outcome R when presented with C on its own, a phenomenon aptly labeled *base-rate neglect*. This "irrational" overreliance on an irrelevant cue differs considerably from the present situation, in which not only $P(F|C) = P(R|C)$ but also $P(C|F) = P(C|R)$—where R and F denote the two equally frequent categories. Another difference between our studies and previous work on base-rate neglect is that in the latter case, people were shown to rely on the irrelevant when it was presented in isolation at test. In the present experiments, by contrast, people used context when it co-occurred with another relevant cue at test.

In a further exploration of irrelevant-cue use, Kruschke (1996, Experiment 4) showed that when a cue was strongly and equally associated with two outcomes, presentation of the cue on its own would elicit the more common outcome. That is, although $P(C|F) = P(C|R)$, people chose outcome F in response to C more frequently than normatively mandated, thus exhibiting an exaggerated reliance on base rates that complements the base-rate neglect observed by Gluck and others.

RASHNL was able to accommodate the results of both Gluck and Bower (1988) and Kruschke (1996), suggesting that reliance on an irrelevant cue can be predicted by the associative learning and attention-shift mechanisms embodied in the theory. Our results, by contrast, show that those mechanisms are insufficient to account for the "irrational" use of an irrelevant cue in KP.

Exemplars Versus Rules

The issue of when people rely on exemplar-based processing and when they use rules or other abstractions has been an important subject of debate (e.g., Erickson & Kruschke, 1998, 2002; Nosofsky & Johansen, 2000; Rouder & Ratcliff, 2004). In previous examinations of probabilistic categorization, exemplar representations have been shown to underlie behavior when the stimuli were few and distinct, whereas any manipulation that rendered stimuli more confusable engendered more rule-based responding (Rouder & Ratcliff, 2004). Our studies, by contrast, demonstrate that people use both types of representations, and to a roughly equal extent, irrespective of the stimuli's distinctiveness (i.e., shading vs. numerosity).

It is noteworthy that we observed clear evidence for partitioning—and hence rule use—in Experiment 2, whose training stimuli were maximally discriminable. By contrast, in Rouder and Ratcliff's (2004) Experiment 3, which also involved highly discriminable stimuli, most participants relied on an exemplar representation. It follows that stimulus discriminability is not the sole determinant of rule use in probabilistic categorization; people use rules even if they require the detection of a subtle relationship between an otherwise irrelevant cue and the relationship among cues along a relevant dimension. Furthermore, rule use in KP did not take the form of a unitary rule representation or a rule-plus-exception representation, but instead was characterized by multiple partial rules whose access was gated by an irrelevant cue. Previous work in artificial grammar learning has demonstrated that people can selectively access different types of knowledge when instructed to do so (see e.g., Brooks, 1978; Dienes, Altmann, Kwan, & Goode, 1995), but in the current tasks, the utilization of the irrelevant context cue to gate rule use emerged without (a) explicit

pointers to different rules or (b) knowledge that there were multiple ways to approach the task.

Why, then, do people partition their knowledge and rely on partial rules? Furthermore, why is partitioning related to intelligence and working memory? Why are people more likely to partition—and hence use rules—if their ability is lower or, equivalently, if the task is complex rather than trivially simple? It is intriguing that previous discussions of the factors underlying rule use have taken opposing views: On the one hand, Rouder and Ratcliff (2004) associated rule use with complexity (i.e., tasks involving many stimuli or stimuli that are difficult to discriminate). On the other hand, Erickson and Kruschke (2002) argued that simple category structures (i.e., those that can be divided by a linear boundary) invite rule use, whereas complex representations are more likely to rely on exemplars. Those opposing suggestions can be reconciled by considering some of the differences between the relevant studies.

There is theoretical (e.g., Feldman, 2006; RULe-plus-EXception model [RULEX]: Nosofsky et al., 1994; Supervised and Unsupervised STratified Adaptive Incremental Network [SUSTAIN]: Love et al., 2004) and empirical (e.g., Johansen & Palmeri, 2002) support for the idea that people first seek simple rules to perform a task and then augment them, during further training, by an exemplar representation as needed. This notion has two implications in the present context. First, given that people in the studies by Rouder and Ratcliff (2004) were trained extensively (i.e., for a minimum of some 2,000 trials), even the discriminable stimuli may have involved early rule use, although this escaped detection because the analysis only considered final performance. Second, in Rouder and Ratcliff's experiments, all stimuli were unidimensional, and the use of rules thus did not permit reduction of dimensionality. This stands in contrast to the two-dimensional spaces used by Erickson and Kruschke (1998, 2002), whose effective dimensionality could be reduced by placing a boundary orthogonal to one of the dimensions, thus permitting rapid error reduction. It follows that if rules are easy to implement (e.g., when the rule boundary is orthogonal to a stimulus dimension or if the objects classified by the rule share a common property; Pothos, 2005) and substantially reduce error, their use may persist irrespective of stimulus discriminability and irrespective of further training, thus apparently linking rules to "simple" tasks. Conversely, if rules cannot substantially and immediately reduce error, as in the unidimensional tasks used by Rouder and Ratcliff (2004), or if the rules are complex, as in the Type IV–VI category problem of Shepard, Hovland, and Jenkins (1961), the incentive persists to create an exemplar representation with additional training. Not surprisingly, this option is facilitated by better stimulus discriminability, thus linking rule use in Rouder and Ratcliff's task with "complexity."

We argue that the present experiments present an instance in which (partitioned) rules permit rapid error reduction, similar to the studies by Erickson and Kruschke (e.g., 2002). This argument is supported by an analysis of complexity provided by Feldman (2006; see also Feldman, 2000, 2003a, 2003b). The unpartitioned category space in the present experiments entails considerable complexity because the values of one stimulus dimension (i.e., context) constrain the values of the other dimension. Interdimensional constraints of this type have been formally identified as sources of complexity (Feldman, 2006). In the extreme case, when all of the dimensional values are constrained by the other dimensional values, a concept cannot be learned by a simple rule but is isomorphic to a set of individual exemplars (i.e., Shepard et al.'s, 1961, Type VI problem). Conversely, when the values of other dimensions are irrelevant to classification on the basis of a given dimension, as in Shepard et al.'s (1961) Type I problem, complexity is minimal (Feldman, 2006). Partitioning reduces the complexity of the rules needed to summarize the categories by removing the constraints placed on the relevant dimension—that is, when the category space is partitioned by context, all stimuli within each context are unidimensional and can be summarized by a simple rule. Furthermore, partitioning obviates the need to develop exemplar representations as per our argument above. In consequence, neither the number of stimuli nor their discriminability affects the prevalence of KP. This argument can be buttressed by considering related findings concerning relational similarity.

Goldstone, Medin, and Gentner (1991) provided a demonstration of people's ability to engage in relational processing of multiple features. Participants were asked to judge whether a target (T) resembled one or the other comparison stimulus (A or B) more. For example, when the target consisted of the three features "× ▮ ×", and the comparison stimuli were "× ○ □" (A) and "□ ○ □" (B), people judged the target to be more similar to B, although B—unlike A—shared none of its features with T. Instead, people chose B on the basis of its relational similarity (two identical features bracketing a different feature). Relational similarity has been identified as being more important than simple feature overlap if it helps identify smaller objects sets, thus simplifying the representation (see, e.g., Tenenbaum & Griffiths, 2001). Likewise, in our experiments, context identified two subsets of stimuli, in each of which the mapping between the ordinality of shading (or numerosity) and the target probabilities was consistent. Thus, partitioning permitted people to exploit the relational similarity among instances within each context.

The preceding analyses of rule-use and simplicity may explain why people of differing ability form different representational solutions to the same problem. The analyses are consonant with the idea that an exemplar representation is only available if one's ability or working memory span can handle the increased complexity. By contrast, a rule-based or heuristic approach offers an equally valid solution (in terms of the task presented during training) without the constraints added by increasing complexity and is thus utilized more often by persons with lower ability.

Toward a Mixture-of-Experts Model of Probabilistic Categorization

Given that neither an exemplar model nor a rule-based approach could, on its own, explain both types of behavior observed in the present studies, a mixture-of-experts approach (e.g., Erickson & Kruschke, 1998; Jacobs, Jordan, Nowlan, & Hinton, 1991) constitutes an obvious avenue for further exploration. Erickson and Kruschke (1998) presented a mixture-of-experts model for category learning, ATRIUM, which combined an exemplar module—instantiated by the ALCOVE architecture that also underlies RASHNL—with one or more rule modules. A rule in ATRIUM consists of a sigmoid boundary that divides a stimulus dimension. Hence, as in GRT, stimuli close to the boundary will be assigned to either category with nearly equal probability, whereas stimuli

that are increasingly distant from the boundary will be assigned to the appropriate category with increasing probability. (The rules in ATRIUM are arguably isomorphic to those in GRT, the primary difference being that the functionality of the perceptual noise in GRT is replaced by the sigmoid shape of the boundary in ATRIUM.) Crucially, unlike GRT, the rules in ATRIUM are learned and thus unequivocally encompass only the information available during training.

ATRIUM has been successfully applied to numerous phenomena, including KP in deterministic categorization (Yang & Lewandowsky, 2004). Given that separate approximations of its constituent modules (i.e., RASHNL and GRT) were successfully applied to the present data, it may appear plausible to expect ATRIUM also to account for our observed KP. However, because the learning mechanisms in ATRIUM are based on those in ALCOVE (Kruschke, 1992), which, as noted earlier, is extremely sensitive to changes in reinforcement schedule (Lewandowsky, 1995), ATRIUM may encounter similar difficulties.

We conclude that a mixture-of-experts model for our results would need to combine an exemplar module with multiple rule modules, all of which must be capable of learning under probabilistic reinforcement (perhaps by including an attenuation of learning rates, similar to RASHNL). To date, no such mixture-of-experts approach exists. We propose that future modeling should build on ATRIUM and related precedents in function learning (Populations Of Linear Experts [POLE]; Kalish et al., 2004).

CONCLUSION

Probabilistic categorization differs from deterministic categorization in a number of important respects (e.g., error is unavoidable in probabilistic tasks, learning rates are decreased when probabilistic reinforcement is introduced, and probabilistic reinforcement tends to disrupt the transfer of rules to new stimuli); despite these differences, we consistently found that some people partitioned the probabilistic learning task, whereas others did not. Those two modes of responding mirrored those found in deterministic categorization. Computational modeling confirmed that when people partitioned, their performance was best captured by a rule-based model, whereas when people did not partition, their performance was best captured by an exemplar model. The current experiments, thus, provide further insight into when people use rules and when people use exemplar and, consistent with our finding that rule use and KP are negatively correlated with intelligence, postulate the reduction of complexity as a determining factor.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723.

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

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition, 31,* 1114–1125.

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., & Lee, W. W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General, 120,* 150–172.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37,* 372–400.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93,* 154–179.

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

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436.

Brooks, L. (1978). Nonanalytic concept formation and memory for instance. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Erlbaum.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information–theoretic approach.* New York: Springer-Verlag.

Busemeyer, J. R., & Myung, I. J. (1988). A new method for investigating prototype learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 3–11.

Cobos, P. L., López, F. J., Rando, M. A., Fernández, P., & Almaraz, J. (1993). Connectionism and probability judgment: Suggestions on biases. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 342–346). Hillsdale, NJ: Erlbaum.

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory & Cognition, 29,* 1165–1175.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 968–986.

Dienes, Z., Altmann, G. T. M., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1322–1338.

Edgell, S. E. (1995). Using configural and dimensional information. In N. J. Castellan (Ed.), *Individual and group decision making: Current issues* (pp. 43–64). Hillsdale, NJ: Erlbaum.

Edgell, S. E., Castellan, N. J., Roe, R. M., Barnes, J. M., Ng, P. C., Bright, R. D., et al. (1996). Irrelevant information in probabilistic categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1463–1481.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review, 112,* 912–931.

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

Erickson, M. A., & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review, 9,* 160–168.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review, 83,* 37–64.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage–retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 556–576.

Farrell, S., & Lewandowsky, S. (2004). Modeling transposition latencies: Constraints for theories of serial order memory. *Journal of Memory and Language, 51,* 115–135.

Feldman, J. (2000, October 5). Minimization of Boolean complexity in human concept learning. *Nature, 407,* 630–633.

Feldman, J. (2003a). A catalog of Boolean concepts. *Journal of Mathematical Psychology, 47,* 75–89.

Feldman, J. (2003b). The simplicity principle in human concept learning. *Current Directions in Psychological Science, 12,* 227–232.

Feldman, J. (2006). An algebra of human concept learning. *Journal of Mathematical Psychology, 50,* 339–368.

Friedman, D., & Massaro, D. W. (1998). Understanding variability in binary and continuous choice. *Psychonomic Bulletin & Review, 5,* 370–389.

Gignac, G., & Vernon, P. A. (2003). Digit Symbol rotation: A more g-loaded version of the traditional Digit Symbol subtest. *Intelligence, 31,* 1–8.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology, 23,* 222–264.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3,* 79–87.

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

Jones, M., Maddox, W. T., & Love, B. C. (2006). The role of similarity in generalization. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (pp. 405–410). Mahwah, NJ: Erlbaum.

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General, 132,* 133–156.

Kalish, M. L., & Kruschke, J. K. (1997). Decision boundaries in one-dimensional categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1362–1377.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review, 111,* 1072–1099.

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

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

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

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review, 12,* 605–621.

Lewandowsky, S. (1995). Base-rate neglect in ALCOVE: A critical re-evaluation. *Psychological Review, 102,* 185–191.

Lewandowsky, S., & Brown, G. D. A. (2005). Serial recall and presentation schedule: A micro-analysis of local distinctiveness. *Memory, 13,* 283–292.

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.

Lewandowsky, S., & Kirsner, K. (2000). Knowledge partitioning: Context-dependent use of expertise. *Memory & Cognition, 28,* 295–305.

Lewandowsky, S., Roberts, L., & Yang, L.-X. (2006). Knowledge partitioning in categorization: Boundary conditions. *Memory & Cognition, 34,* 1676–1688.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111,* 309–332.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics, 53,* 49–70.

Maddox, W. T., & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes, 66,* 309–332.

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural learning-based system in category learning. *Psychonomic Bulletin & Review, 11,* 945–952.

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.

McKinley, S. C., & Nosofsky, R. M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 294–317.

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

Mehta, R., & Williams, D. A. (2002). Elemental and configural processing of novel cues in deterministic and probabilistic task. *Learning and Motivation, 33,* 456–484.

Myers, J. L. (1976). Probability learning and sequence learning. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Approaches to human learning and motivation* (pp. 171–205). Hillsdale, NJ: Erlbaum.

Myers, J. L., Lohmeier, J. H., & Well, A. D. (1994). Modeling probabilistic categorization data: Exemplar memory and connectionist nets. *Psychological Science, 5,* 83–89.

Navarro, D. J. (2007). On the interaction between exemplar-based concepts and a response scaling process. *Journal of Mathematical Psychology, 51,* 85–98.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences with Dirichlet processes. *Journal of Mathematical Psychology, 50,* 101–122.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Nosofsky, R. M. (1998). Selective attention and the formation of linear decision boundaries: Reply to Maddox and Ashby (1998). *Journal of Experimental Psychology: Human Perception and Performance, 24,* 322–339.

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 282–304.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review, 7,* 375–402.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211–233.

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review, 5,* 345–369.

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

Nosofsky, R. M., & Stanton, R. D. (2005). Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 608–629.

Oberauer, K., Schulze, R., Wilhelm, O., & Süβ, H.-M. (2005). Working memory and intelligence—Their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*(1), 61–65.

Pelli, D. G. (1991). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442.

Pothos, E. M. (2005). The rules versus similarity distinction. *Behavioral and Brain Sciences, 28,* 1–49.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Comparing connectionist and diffusion models of reaction time. *Psychological Review, 106,* 261–300.

Rouder, J. N., & Ratcliff, R. (2004). Comparing categorization models. *Journal of Experimental Psychology: General, 133,* 63–82.

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology, 42A,* 209–237.

Shanks, D. R. (1991). A connectionist account of base-rate biases in categorization. *Connection Science, 3,* 143–162.

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior and Processes, 24,* 405–415.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making, 15,* 233–250.

Shepard, R. N. (1987, September 11). Toward a universal law of generalization for psychological science. *Science, 237,* 1317–1323.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75,* 1–45.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–640.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14,* 101–118.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11,* 192–196.

Yamauchi, T., & Markman, A. B. (2000). Learning categories composed of varying instances: The effect of classification, inference, and structural alignment. *Memory & Cognition, 28,* 64–78.

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

Yang, L.-X., & Lewandowsky, S. (2004). Knowledge partitioning in categorization: Constraints on exemplar models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1045–1064.

Yang, L.-X., & Lewandowsky, S. (2005). *Spontaneous knowledge partitioning in categorization.* Paper presented at the 46th Annual Meeting of the Psychonomic Society, Toronto, Ontario, Canada.

Yang, L.-X., Lewandowsky, S., & Jheng, W. (2006, July). *Working memory and knowledge partitioning in categorization.* Paper presented at the 4th International Conference on Memory, Sydney, Australia.

Young, M. E., Wasserman, E. A., Johnson, J. L., & Jones, F. L. (2000). Positive and negative patterning in human causal learning. *Quarterly Journal of Experimental Psychology, 53B,* 121–138.

Received June 22, 2007
Revision received March 5, 2008
Accepted March 10, 2008 ■