Sensitivity to Correlation in Probabilistic Environments

Daniel R. Little
Bachelor of Arts (Honours in Psychology)

This thesis is presented for the degree of Doctor of Philosophy
of The University of Western Australia
School of Psychology

2008
“There is much pleasure to be gained from useless knowledge.”

Bertrand Russell

“A man’s errors are his portals of discovery.”

James Joyce
Abstract

Natural categories seem to be comprised of clustered stimuli that contain a myriad of correlated features; birds, for example, tend to fly, have wings, lay eggs, and make nests. Nonetheless, the evidence that people use these correlations during intentional category learning is overwhelmingly negative (Murphy, 2002). People do, however, show evidence of correlational sensitivity during other types of category learning tasks (e.g., feature prediction). The usual explanation is that intentional category learning tasks promote rule use, which discards the correlated feature information; whereas, other types of category learning tasks promote exemplar storage, which preserves correlated feature information. However, all of the intentional category learning tasks employed to examine correlational sensitivity to date have only used deterministic mappings of stimuli to categories (i.e., each stimulus belongs to only one category). The current thesis is concerned primarily with the effects introducing the probabilistic assignment of stimuli to categories on the acquisition of different types of correlational knowledge. If correlational knowledge depends on whether or not people selectively attend to the correlation then probabilistic reinforcement, which is predicted to increase attention shifting (Kruschke & Johansen, 1999), should lead to increased correlational sensitivity. The first paper of this thesis confirms that selective attention provides a way to explain the presence or absence of correlational knowledge in different tasks. However, selective attention models have been unable to account for tasks in which people use the correlation between a non-relevant cue and regions of the category space to switch between the application of multiple rules. This phenomenon, known as knowledge partitioning, is explored in the second paper of this thesis. This thesis
also extends the empirical implications of the first two papers to existing research (see included paper 3) and also provides recommendations of how utilize this conceptualization of knowledge for practitioners in the applied setting (see included paper 4). Finally, in addition to increasing attention shifting, probabilistic feedback is also assumed to result in an attenuation of learning over time (Kruschke & Johansen, 1999); the fifth paper in this thesis provides empirical confirmation that people attenuate learning in response to unavoidable error.
Acknowledgments

I would like give my sincere gratitude to my PhD supervisor, Professor Stephan Lewandowsky; his guidance and enthusiasm for my work made the process of completing this thesis much easier. He also provided the impetus for many epiphanies that made this thesis possible and made my world a richer place. The members of my review panel provided many useful ideas which guided this work; thank you to John Dunn, Gilles Gignac and Murray Maybery. I would be remiss not to acknowledge the members of the Cognitive Science Laboratories, whose steadfast sense of humor and camaraderie made this experience most enjoyable. Special thanks to Leo and Dave. Thanks to Stewart, Lee-Xieng, and Joe. I am grateful for Red Music and Vincent Price for No-Fi and for granting an outlet for my other creative ideas; thanks to Heath and Ben. I would like to thank Dean, Terry and Sarah Little and Lydia Hart for their support. Finally, thank you to Jade for teaching me to challenge what I think.
# Contents

1 Introduction 3

2 Category learning and probabilistic environments 5
   2.1 Concept Usage Tasks 6
   2.2 Concept Learning Tasks 7
      2.2.1 Deterministic Category Learning 9
      2.2.2 Function Learning 11
      2.2.3 Multiple-cue Probability Learning 12
      2.2.4 Probabilistic Categorization 13

3 Selective Attention 15
   3.1 Selective Attention in Deterministic Category Learning 16
   3.2 Selective Attention in Sensitivity to Correlation 17

4 Knowledge Partitioning 19
   4.1 Knowledge Partitioning and Correlated Cues 20
   4.2 Knowledge Partitioning and Selective Attention 22

5 Overview of Included Papers 25
   5.1 Correlated Cues and Probabilistic Feedback 25
      5.1.1 Summary of Method 26
      5.1.2 Summary of Results 26
      5.1.3 Summary of Discussion 27
Papers Included in this Thesis

Paper 1 (Refereed)


Paper 2 (Refereed)


Paper 3


Paper 4 (Refereed)

Paper 5 (Refereed)

Contribution of Candidate to Submitted Work

**Paper 1** 80% contribution.
Completed all experimental design, programming, experimental testing, data analysis, computational modeling, and manuscript preparation. Final manuscript revision was aided by a significant contribution from Stephan Lewandowsky, who also supervised the project. This paper was recently submitted to the *Journal of Experimental Psychology: Learning, Memory and Cognition*.

**Paper 2** 80% contribution.
Completed all experimental testing, data analysis, computational modeling, and manuscript preparation. The majority of the manuscript revision was undertaken by myself with a significant contribution from Stephan Lewandowsky, who also supervised the project and provided the initial experimental design. This paper was well received during its first submission to the *Journal of Experimental Psychology: Human Perception and Performance* and is currently in press in that journal.

**Paper 3** 100% contribution
Sole author.
Paper 4  30% contribution

Wrote the section on transfer and the final conclusions.

Paper 5  15% contribution

Programmed the initial model used in the paper, completed pilot testing of the experimental idea, provided experimental code, conducted the regression analysis and GCM modeling reported in the paper and made a significant minor contribution to manuscript revision. This paper was well received at its first submission to Memory & Cognition; the paper was revised and resubmitted and is currently undergoing its second review.
List of Tables

11.1 Complexity of the category representations shown in Figure 11.1. . . . . . 313
List of Figures

2.1 Three-dimensional classification of concept learning paradigms .......................... 8

3.1 Attention in psychological space from Nosofsky (1986) ....................................... 16

4.1 Stimulus space from Yang & Lewandowsky (2003). Training items are represented by circles (Category A) and crosses (Category B). Transfer items are represented by filled diamonds. The category boundary is represented by the solid line; all stimuli above the solid line belong to Category B, whereas all stimuli below the line belong to Category A. .................................................... 21

11.1 Data and theoretical context insensitive (CI) and knowledge partitioning (KP) hypotheses (i.e., representations) ................................................................. 315

12.1 Example cuboid category space and cuboid hypotheses for complexity analysis. 329
Chapter 1

Introduction

Environmental constraints and natural laws necessarily imply that the world has a definite structure. For example, the seasons follow a cycle that co-occurs with particular calendar months due to, among other things, the interaction of the Earth’s revolution around the Sun and the tilt of the Earth’s axis. In a similar fashion, small birds, such as the Japanese quail, tend to sing more in the spring as the change in the amount of sunlight triggers hormonal changes which lead to increased mate attraction behaviors (Yoshimura et al., 2003). To date, however, it is unclear whether or not a human learner set to work learning this structure acquires all of the information that is available to learn. For instance, a naive observer who set about describing the mating behavior of birds might not ever learn that small birds tend to sing more than large birds. Most category learning research to date suggests that people tend to acquire whatever knowledge is relevant for the task at hand, but is this conclusion the result of learning in a perfect (i.e., deterministic) environment, where the same stimuli always belong to the same categories and perfect rules can be developed to guide classification? The current thesis explores this question by comparing categorization behavior in perfect, deterministic environments with behavior in imperfect, probabilistic environments.

This thesis adduces that selective attention, which has successfully been used to explain several benchmark category learning phenomena, plays a central role in whether or not
structural information which is not relevant to the goal of the task (instantiated here as the correlation between cues) is acquired. To that end, how attention interacts with the structure of probabilistic environments to produce correlational sensitivity and the limits of the explanatory capabilities of selective attention in probabilistic environments are examined. This thesis also examines how selective attention can explain correlational sensitivity outside of category learning, and tests one of the corollary assumptions of a successful selective attention model of probabilistic category learning.

This thesis is organized as a collection of journal articles, and therefore, inevitably follows a different structure and style than a traditional doctoral dissertation. Each included paper is self-contained with its own introduction, conclusions and implications. Hence, the organization of this thesis is as follows. The scope and aims of the thesis are provided in Chapters 2, 3, and 4. In Chapter 2, the particular research environment is outlined in moderate detail to frame the work that follows. Namely, probabilistic category learning is placed in context within the broader concept learning paradigm, and additionally, justification is provided for how correlations between features can stand as a proxy for the structure in the world. Chapter 3 provides a brief overview of selective attention with particular emphasis on how it may interact with probabilistic feedback and correlated cues, and Chapter 4 briefly describes the knowledge partitioning paradigm with particular emphasis on how knowledge partitioning relates to correlated cue and selective attention research. Chapter 5 is an overview of the included papers that describes how they fit within the framework established in the preceding chapters. Chapters 6 through 10 are the included papers and Chapter 11 briefly summarizes the overall findings of the thesis.
Chapter 2

Category learning and probabilistic environments

Whether or not people are sensitive to the correlational structure of the environment is important to ascertain for several reasons. Firstly, natural categories are partly defined by their correlated features. For example, all members of the taxonomic category 'birds' would tend to have feathers, be able to fly, lay eggs, and so on. As noted by Murphy (2002), the prototypical example of a concept could be defined by being comprised of all of the correlated features for that concept. Knowledge of category correlations would allow for more accurate predictions to be made about future encounters with members of the category. Hence, having and making use of correlational knowledge is rational (see e.g., Anderson, 1991a, 1991b). Secondly, the correlational structure of the concept identifies particular relations which demarcate particular category subgroups (Chin-Parker & Ross, 2002); hence, correlational knowledge would simplify movement between different levels of a category hierarchy (i.e., basic level, subordinate level and superordinate level; see e.g., Rosch, 1978; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Finally, further examination of correlational knowledge in category learning is necessary because the evidence for the existence of correlational knowledge is mixed.

Whether or not people show evidence of correlational sensitivity depends partly on
the goal of the current learning task (McNorgan, Kotack, Meehan, & McRae, 2007). If the goal is to learn the category by making deliberate category predictions based on presentations of members of that category then people do not appear to have access correlational information. By contrast, if the goal is to use the category members to achieve some auxiliary goal (i.e., like predicting missing features; Chin-Parker & Ross, 2002; or making typicality ratings; Ahn, Marsh, Luhmann, & Lee, 2002), then people appear to have access to the correlational information as a by-product of interacting with the categories.

2.1 Concept Usage Tasks

There are a broad array of tasks which fall under the rubric of concept usage. For instance, feature inference tasks require participants to use knowledge of the category and any given features to predict missing features (Yamauchi, 2000, 1998). When trained to predict missing features from a category space which contains correlated features, participants readily learn the correlation (Chin-Parker & Ross, 2002; Crawford, Huttenlocher, & Hedges, 2006). This is perhaps unsurprising because the task involves active interaction with the features; hence, knowledge of the correlation between the features is necessary to perform the task (e.g., if you know that a bird is small, you can correctly predict the missing characteristic of singing). In incidental category tasks, participants are required to use category knowledge to rate stimuli on their typicality (Ahn et al., 2002; McNorgan et al., 2007) or on 'likability' (Wattenmaker, 1991, 1993). In incidental tasks, participants again show sensitivity to correlational structure. Usually in these tasks natural stimuli are used and prior knowledge of the category domain can be brought to bear on the task; importantly, participants tend to only show sensitivity to correlations which are consistent with or expected by prior knowledge (cf. Murphy & Wisniewski, 1989). Although sensitivity

---

1The term “correlation” has been used to define a wide array of different stimulus relationships. Throughout most of this thesis, the term ‘correlation’ will be used to refer only to feature co-occurrences that are not diagnostic of category membership. Hence, knowledge of category structure will be assessed independently of the diagnostic elements that demarcate the category bounds. Where other definitions are used, they will be made clear by the context of the text. Other definitions of ‘correlation’ are explored in Chapter 8.
to correlation appears to be evident in the concept usage paradigms, the evidence from concept learning tasks is less clear.

2.2 Concept Learning Tasks

Like concept usage, the term 'concept learning' subsumes a wide array of tasks; hence, we find it helpful to differentiate these tasks by creating a classification of paradigms with respect to three principal dimensions: The nature of the cues (discrete vs. continuous); the nature of the response (discrete vs. continuous); and the type of feedback provided during learning (deterministic vs. probabilistic). For present purposes, any cue or response that is binary or involves a nominal scale (e.g., "red" vs. "green" vs. "blue" or "category A" vs. "category B") is considered discrete; conversely, any cue or response that is expressed on at least an ordinal scale (e.g., five levels of increasing saturation of a single color) is considered continuous. In terms of stimulus feedback, deterministic feedback refers to situations where items are always reinforced to the same category, while probabilistic feedback refers to situations when the relationship between the item and the category is determined by some probability less than one. As shown in Figure 2.1, there are three main types of paradigms that require consideration: 1) category learning with deterministic feedback, 2) function learning, and 3) category learning with probabilistic feedback. Sensitivity to correlation has been examined in each of these paradigms in different ways.
<table>
<thead>
<tr>
<th>Feedback</th>
<th>Deterministic</th>
<th>Probabilistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Discrete</td>
<td>Continuous</td>
</tr>
<tr>
<td>Categorization</td>
<td>(e.g., Medin &amp; Shaffer, 1978; Nosofsky, 1984; Shepard, Hovland, &amp; Jenkins, 1961)</td>
<td>Multiple Cue Judgment (e.g., Justlin, Olson, and Olsson, 2003)</td>
</tr>
<tr>
<td>Continuous Cues</td>
<td></td>
<td>Multiple Cue Judgment (e.g., Justlin, Olson et al., 2003)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Function Learning (e.g., Kalish, Lewandowsky, &amp; Kruschke, 2000; Yang &amp; Lewandowsky, 2003)</td>
</tr>
</tbody>
</table>

Figure 2.1: Three-dimensional classification of concept learning paradigms
2.2.1 Deterministic Category Learning

Category learning is typically studied under deterministic feedback conditions (though see e.g., Estes, 1984). Research within the category learning paradigm has focused on the development of computational models of cognitive processes and representations; typically this research only addresses sensitivity to the diagnostic elements of the category structure. That is, category learning models are usually only concerned with how people use the correlation of cues with some criterion or category outcome. For instance, in deterministic category learning tasks, people are adept at learning which stimulus features are diagnostic, that is, which features are useful for predicting the correct category outcome (Shepard, Hovland, & Jenkins, 1961; Medin & Schaffer, 1978; Nosofsky, 1986; Kruschke, 1992). Furthermore, if a correlation is diagnostic then people are sensitive to it. For instance, Wattenmaker (1993) tested participants ability to extract coherent properties from a single category with two perfectly co-occurring cues (i.e., the values of the cues were always the same); that is, participants had to determine what made the category cohere as a category. At a later test, participants demonstrated sensitivity to the correlated stimulus dimensions by selecting new stimuli which preserved the correlation shown at training as more typical of the studied category than new stimuli which broke the correlation. However, learning a single category with two perfectly co-occurring cues is equivalent to learning a bi-conditional rule (i.e., all non-members of the category will have opposing values on the correlated dimensions). Hence, participants in this experiment might have been sensitive to the fact the co-occurring features perfectly determined the category. Extracting information about whether people are sensitive to category structure is difficult because correlated features are often confounded with diagnosticity (Medin, Altom, Edelson, & Freko, 1982; Wattenmaker, 1993).

Several studies have explicitly studied sensitivity to non-diagnostic correlated features, and the usual result is that people are not sensitive to non-diagnostic structure in deterministic category learning (Chin-Parker & Ross, 2002; Wattenmaker, 1991). For example, Wattenmaker (1991) instructed participants to learn as much as possible about members
of a single category that contained several non-diagnostic cue correlations. After training, participants demonstrated little ability to access information about these correlated cues. Instead, participants were aware only of the diagnostic dimensions and configurations. Similar results were found by Chin-Parker and Ross (2002) who tested the same category structure containing non-relevant correlated cues in both a feature inference task and a deterministic category learning task. Correlational sensitivity was evident in the former task but not the latter.

The absence of correlational sensitivity in category learning is an important test of categorization models because some classes of models predict an absence of correlational knowledge whereas others do not. For example, use of a rule during category learning allows one to differentiate members of a category (i.e., you would be able to identify which animals are birds and which are not) but predicts that no information about the relationships among the features should be available (e.g., that small birds are more likely to sing than large birds). By contrast, because exemplar models store information about all features of each particular instance, correlational information would be predicted to be available (i.e., information about singing and bird size could be extracted from the stored exemplars).

In summary, most deterministic category learning research demonstrates that people learn diagnostic features but not nondiagnostic correlational features. The experiments in the current thesis use a category learning task but with probabilistic feedback. In probabilistic category learning, although individual responses are necessarily discrete rather than continuous, people 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 (Myers, 1976; Vulkan, 2000). The observed aggregate response probabilities across trials can therefore arguably be interpreted as a response "magnitude", akin to the continuous magnitudes expressed on each trial during function learning. If the probability with which a cue predicts the target outcome is taken to represent a continuous target magnitude, then a category learning task with probabilistic feedback can be considered an
analogue of the function learning task (DeLosh, Busemeyer, & McDaniel, 1997; Kalish, Lewandowsky, & Kruschke, 2004; Lewandowsky, Kalish, & Ngang, 2002); hence, it is worthwhile to consider correlational sensitivity in this domain.

2.2.2 Function Learning

Function learning can be considered an extension of category learning, where a continuous response variable rather than a discrete response must be predicted on the basis of a continuous stimulus variable (DeLosh et al., 1997; Kalish et al., 2004; Lewandowsky et al., 2002). For example, people may learn how long to water the lawn as a function of the day’s temperature, or what one’s blood alcohol content will be as a function of the number of cocktails consumed. A defining attribute of function learning is the ability to extrapolate beyond the range of training stimuli: Suppose one has learned that two cocktails raise one’s blood alcohol to .04, whereas three and four cocktails result in .06 and .08, respectively, then one can offer a good guess—at least prior to consumption—as to what would be expected after six cocktails.

Function learning research has typically been concerned with the type of functional forms humans can learn and the relative speed at which they learn them. For instance, learning functions that are created by randomly pairing a stimulus with a response results in poor learning and large response times (Carroll, 1963; Fitts & Deininger, 1954). Functional forms with a more coherent structure, such as positive and negative linear functions or concave and convex parabolic functions are able to be learned (Björkman, 1965; Sheets & Miller, 1974). Of these four function types, linear relations are easier to learn than non-linear relationships (Sheets & Miller, 1974) and a positive relationship is easier to learn than a negative relationship (Björkman, 1965; Brehmer, Kuylentierna, & Liljergren, 1974). Sensitivity to correlation has not been directly examined in function learning, but there have been many studies looking at the effect of redundant (i.e., correlated) cues on judgment in a related domain, multiple-cue probability learning.
2.2.3 Multiple-cue Probability Learning

The multiple-cue probability learning (MCPL) paradigm developed following Brunswik’s (1943) assertion that the probabilistic nature of the environment requires that people (and other organisms) adopt probability-based strategies. While the deterministic category learning paradigm emphasized the development of transfer designs to examine patterns of generalization and computational models to elucidate cognitive processes; research in MCPL focused on mapping the boundary conditions of the type of structures that humans could learn under probabilistic conditions. In MCPL, people are trained to predict a probabilistically reinforced, continuous criterion based on one or more continuous cues.

Naylor and Schenck (1968) tested participants’ sensitivity to correlated cues by training people to predict a continuous cue that varied as a function of two correlated cues. In this experiment and in other MCPL tasks, the cues co-varied along continuous dimensions determined by some underlying correlation. Participants were sensitive to the correlated cues, but the cue correlations were confounded with diagnosticity because both cues were equally valid (i.e., the value of one cue could be predicted from the value of the other but so could the value of the criterion).

Armelius and Armelius (1974) found that participants’ achievement and consistency (i.e., how well participants responses match the actual correct response and the stability of those responses, respectively) decreased as the cue correlation increased; that is, as more redundancy (i.e., more non-relevant correlations) is introduced into the environment the greater the performance decrement. Schmitt and Dudycha (1975) found that when correlations between cues were not confounded with cue validity, performance was worsened as the cue correlation became larger, similar to the effect of introducing an irrelevant cue in category learning (Edgell et al., 1996; Edgell & Hennessey, 1980). Hence, in MCPL, performance does appear to be affected by the presence of correlated cues.

A variant of this task is explored in this thesis by using a discrete criterion instead of a continuous criterion. If a discrete criterion is paired with discrete cues then the task is termed ‘non-metric multiple cue probability learning’ (see Figure 2.1). If continuous
cues are retained but the criterion is discrete then the task is termed 'metric multiple cue probability learning' (see Figure 2.1). In light of the similar results found across discrete and continuous cues in deterministic category learning (Kalish & Kruschke, 2000; Kalish, 2001; Nosofsky & Palmeri, 1998) and to simplify the terminology used in this thesis, both of these tasks (i.e., non-metric and metric multiple cue probability learning) will be referred to as probabilistic categorization.

2.2.4 Probabilistic Categorization

In probabilistic categorization, people 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 category A on 75% of all relevant 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 medical diagnosis to predicting the weather; and accordingly, research in probabilistic categorization has frequently been considered as a tractable approach to studying real-world decision making (Brunswik, 1939; Edgell, 1983, 1980; Edgell & Morrissey, 1992, 1987; Kruschke & Johansen, 1999).

On the one hand, in terms of stimulus presentation parameters, probabilistic categorization is identical to the traditional, deterministic category-learning paradigm, in which a multi-dimensional stimulus is presented for classification into one of several candidate categories, and in which the participant’s response is followed by corrective feedback. Indeed, the surface characteristics of a single probabilistic learning trial resemble those of a deterministic learning trial with the following exceptions: First, the cues in probabilistic categorization need not be intrinsic attributes of the stimulus (e.g., cues may be discrete and either present or absent as in the weather prediction task; see Figure 2.1) in contrast to the dimensions of a categorization stimulus that take on different values. Second, the binary outcome in probabilistic categorization need not consist of conventional categories (e.g., outcomes may be mutually exclusive opposites). Third, unlike most categorization studies, the probabilistic nature of the task implies that perfect performance is usually not
attainable.

Like its continuous counterpart, MCPL, research in probabilistic category learning has focused on testing the boundary conditions of what humans can learn in probabilistic environments with a particular emphasis on how cue utilization is affected by a cue’s validity and salience. Salience refers to a cue’s ability to attract utilization regardless of its validity. For instance, a typical result from this domain is that stimulus cues and compounds are utilized in proportion to their validity (Edgell et al., 1996; Edgell, 1978, 1980). However, cue validities interact to produce cue competition effects such that raising the validity of a previously non-relevant cue decreases the utilization of previously relevant cue (Edgell et al., 1996; Edgell, 1978). Salience has a similar effect on utilization; namely, that a cue of greater salience is utilized more than a cue of lesser salience but equal validity, and validity and salience trade off in a predictable manner (Kruschke & Johansen, 1999). In order to account for the combined effects of validity and salience in probabilistic learning, Kruschke and Johansen (1999) introduced a theory of learning, RASHNL, that proposes that whenever an error is encountered during category learning, attention is rapidly shifted among the stimulus dimensions. This idea accounts for cue interaction effects by assuming that sometimes a less valid cue will ‘steal’ attention from a more valid cue. It is conceivable that rapidly shifting attention might provide a mechanism for explaining the presence or absence of correlational knowledge in different domains. This possibility is outlined in Chapter 3.

Another property of performance in probabilistic domains is that learners tend to decrease their learning rates over time as they become accustomed to an unavoidable level of error (Brunswik, 1939; Kruschke & Johansen, 1999). This phenomenon is explored empirically in Chapter 10.
Selective Attention

Knowledge of correlations would be complete if people utilized all available information. However, during category learning, attention limits the information that is effectively available. The basic idea is that as stimulus dimensions are attended to, the mental representation of that stimulus is stretched in psychological space along the attended dimension (Nosofsky, 1986). Because selective attention is assumed to be of limited capacity, unattended dimensions are 'shrunk' in psychological space. For example, the bottom panel of Figure 3.1 shows the effect of attending to the color dimension of a three dimensional stimulus space. The top panel shows the psychological representation before attention is allocated to a dimension. A consequence of attention is that stimuli become more discriminable on attended dimensions and less discriminable on unattended dimensions. Hence, differences in color for the stimuli shown in Figure 3.1 would become more noticeable than differences in shape or size.

In most models that contain selective attention mechanisms, it is assumed that attention is optimally distributed among features (i.e., more diagnostic stimulus features receive higher attentional weightings than less diagnostic features (Shepard et al., 1961). This is a basic assumption of the Generalized Context Model (GCM; Nosofsky, 1984, 1986, 1987, 1989) and variants of the GCM (e.g., D. W. Aha & Goldstone, 1992; D. Aha & Goldstone, 1990; D. Aha & McNulty, 1989; Lamberts, 1998, 1995; Rodrigues & Murre, 2007) and
optimal attention arises as the asymptotic behavior of a connectionist learning models such as ALCOVE (Kruschke, 1992) and RASHNL (Kruschke & Johansen, 1999). This assumption has successfully accounted for many of the benchmark findings in deterministic category learning; hence, we briefly review these findings and extension to the basic selective attention mechanism before examining how selective attention might enable detection of correlation in probabilistic category learning.

3.1 Selective Attention in Deterministic Category Learning

The idea that dimensions are stretched and shrunk in psychological space was a key component of Nosofsky (1984, 1986) GCM model, which successfully provided a link between identification and categorization by assuming that attention is allocated differently in both tasks. Similarly, selective attention can explain the relative difficulty of learning tasks of different complexity (i.e., when fewer or more cues are valid; Shepard et al., 1961) by assuming that it is easier to attend to a few relevant dimensions rather than many
relevant dimensions (Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994). Given that attention is not only directed toward relevant dimensions but that nonrelevant dimensions are learned to be actively ignored, selective attention can account for the blocking of one cue by another (i.e., when a highly relevant cue learned early during training prevents utilization of another equally relevant but later learned cue; Kruschke & Blair, 2000; Kruschke, 2005, 2006). Importantly, data from eye tracking studies have found that gaze location is highly correlated with the attentional distribution predicted by the selective attention models (Kruschke, Kappenman, & Hetrick, 2005; Rehder & Hoffman, 2005a, 2005b).

Several extensions to basic selective attention mechanisms have been proposed, one of which is the rapidly shifting attention mechanism in RASHNL that has successfully accounted for several key results in probabilistic category learning.

### 3.2 Selective Attention in Sensitivity to Correlation

Attention shifting is rational because it enables rapid error reduction (Kruschke, 2003). For example, persistent use of an irrelevant cue, such as displayed in the functional fixedness effect (Birch & Rabinowitz, 1951; Duncker, 1945), would lead to incorrect problem solutions; a better strategy would be to use a different cue after an error is made until the cue which predicts the correct solution is found. In probabilistic environments attention shifting should be increased compared to deterministic environments because of the unavoidable level of error in probabilistic tasks. Cue competition effects in probabilistic category learning attest to this claim. If attention shifting is increased, then sensitivity to correlation might also be increased. For example, Dieckmann (2005) found that in environments of high cue redundancy, people will selectively utilize the most predictive cue more than in situations of low redundancy; that is, when there are multiple probabilistically predictive cues that point to multiple outcomes, people will search more cues to find the most predictive cue before they make a decision. This echoes the finding that people in deterministic categorization are not sensitive to correlated cues and supports the idea that people in probabilistic
environments will shift attention among multiple cues. Chapter 6 examines this prediction.

To summarize, rapidly shifting selective attention predicts that correlational sensitivity should be increased in probabilistic category learning. The null correlational sensitivity in deterministic category learning can be explained by selective attention to relevant dimensions. Selective attention, however, can not explain the heterogeneous representation of knowledge that is found across a wide spectrum of concept learning tasks. The study of this phenomenon, known as knowledge partitioning, is briefly outlined in the following chapter.
Chapter 4

Knowledge Partitioning

Knowledge partitioning is the idea that knowledge, such as category boundaries or exemplars or possibly the probabilistic mappings between cues and outcomes in probabilistic environments, may be fractionated into independent parcels that are used selectively and without reference to knowledge held in other parcels (Lewandowsky et al., 2002; Lewandowsky & Kirsner, 2000; Yang & Lewandowsky, 2003, 2004). In consequence, people may provide contradictory answers to an identical problem, depending on which knowledge parcel they use to guide their answer. Knowledge partitioning has been shown to arise with experts (Lewandowsky & Kirsner, 2000), with non-expert participants in a function learning paradigm (Lewandowsky et al., 2002), and with non-expert participants in categorization tasks involving numeric (Yang & Lewandowsky, 2003) as well as various perceptual stimuli (Lewandowsky, Roberts, & Yang, 2006; Sewell, 2008; Yang & Lewandowsky, 2004). In each instance, a normatively irrelevant context cue (such as a verbal label or the color of the stimulus) triggered the contradictory resolution of an identical problem.

To illustrate, consider a function-learning study in which participants learned a U-shaped (quadratic) function (Lewandowsky et al., 2002). During training, stimulus values were accompanied by one of two context labels, each of which was preferentially associated with one half of the function (i.e., the ascending or descending component of the quadratic). At transfer, people not only used the context label to guide their extrapolation, but their
performance within each context closely resembled the transfer performance of people in two control conditions who had only learned one half of the function. This pattern implies that when presented with an old stimulus in a new context, participants’ performance was not influenced by their acquired knowledge in the original context, suggesting that people partitioned their knowledge into independent parcels.

Yang and Lewandowsky (2003) extended the knowledge partitioning framework to deterministic categorization by accompanying stimuli with a context cue that by itself did not predict category membership but predicted which of two partial boundaries would correctly classify a stimulus. At transfer, participants who partitioned their knowledge were found to rely exclusively on the boundary identified by the stimulus context. Yang and Lewandowsky (2004) additionally showed that while an exemplar model (ALCOVE; Kruschke, 1992) could account for the data from participants who ignored context and used both partial boundaries simultaneously, the model was unable to account for participants who partitioned their knowledge. By contrast, a hybrid model that combined exemplar- and rule-based representations (ATRIUM; Erickson & Kruschke, 1998), successfully captured the performance of participants who partitioned their knowledge.

### 4.1 Knowledge Partitioning and Correlated Cues

Partitioning the task into smaller components requires utilization of the correlation between the context cue and the remaining cues. The context cue is typically a non-relevant binary cue, the values of which are mapped to different areas of the stimulus space where different rules can be used to categorize the stimuli. To illustrate, consider the category space used by Yang and Lewandowsky (2003) shown in Figure 4.1.

In this study, participants were trained to categorize numerical stimuli defined by two dimensions. A third dimension, not shown in Figure 4.1, was added by presenting the stimuli in two different contexts, which were instantiated as two arbitrary labels. All of the training instances were clustered around the local diagonals of the true boundary, which is represented by the solid line in Figure 4.1 given by $Y = 500 - |X - 400|$. Participants
Figure 4.1: Stimulus space from Yang & Lewandowsky (2003). Training items are represented by circles (Category A) and crosses (Category B). Transfer items are represented by filled diamonds. The category boundary is represented by the solid line; all stimuli above the solid line belong to Category B, whereas all stimuli below the line belong to Category A.

were trained that anything above the boundary belonged to Category A and anything below the boundary belonged to Category B. Context was systematically paired with the values of dimension $X$, such that one context (C1) was always shown when $X < 400$ and another (C2) was shown when $X > 400$. The results indicated that some participants had formed two extended partial boundaries (indicated by the dotted lines in Figure 1) rather than responding based on the true boundary. Furthermore, these two partial boundaries were linked with context. For example, in the first context (C1), people would use one of the diagonal boundaries, and in context C2, people would use the other diagonal boundary. Contradictory responding arose in Areas 2 and 4 for the same items presented in different contexts (see Figure 4.1). Hence, it is the utilization of the correlation between
the non-relevant context cue and a particular rule boundary found in part of the category space that characterizes knowledge partitioning. ¹

4.2 Knowledge Partitioning and Selective Attention

Unlike the correlated cues that are explored in Chapter 6, the correlation in knowledge partitioning provides a distinct challenge to selective attention models of category learning. Yang and Lewandowsky (2004) found that an exemplar model with selective attention (i.e., ALCOVE; Kruschke, 1992) could not demonstrate the contradictory responding demonstrated by the knowledge partitioning participants. Only when selective attention was paired with a representational gating mechanism which selected amongst predetermined rule modules (i.e., ATRIUM; Erickson & Kruschke, 1998) could the knowledge partitioning results be explained. Similar mixture-of-expert mechanisms have been successful at explaining similar behavior in function learning tasks (Kalish et al., 2004).

What these results suggest is that people are attending to stimuli in a manner that goes beyond simple selective attention to diagnostic (or salient) dimensions. One criticism of exemplar models (such as the GCM; Nosofsky, 1986) is the assumption that attention is distributed in a way that optimizes performance. An alternative proposal is that attention is distributed in a manner which simplifies a complex problem. For instance, in unsupervised category learning, participants generally prefer to create less complex categories (Colreavy & Lewandowsky, in press; Pothos & Chater, 2002). The unsupervised learning and knowledge partitioning results suggest that when the goal of accurate performance is removed people will opt for simplification. Hence, in knowledge partitioning experiments, where partitioning the category space does not lead to a decrement in performance, partitioning might provide a method of simplifying a complex problem by breaking it into simpler parts. This thesis provides a quantitative analysis of the complexity of knowledge partitioning representations;

¹Though the categories studied in knowledge partitioning studies have been artificial, natural categories also show the same basic trend of having a different rule being marked by a non-relevant cue. One example of this effect using natural concepts comes from the subtyping literature, where stereotypes are often maintained by using a non-relevant but salient cue to differentiate exceptions from rule-consistent stimuli (see e.g., Bott & Murphy, 2007).
though to present this analysis in the context of the experimental findings it is deferred to
the general discussion in Chapter 11.

In summary, knowledge partitioning involves knowledge of the correlation between
a non-relevant cue and different areas of the stimulus space. Simple attention to the
non-relevant context cue can not explain the differential application of rules in opposing
contexts. Whether this phenomenon holds in probabilistic category learning is explored
in Chapter 7. A brief overview of each of the papers is presented before considering the
findings of the research in this thesis.
Chapter 5

Overview of Included Papers

This chapter provides an executive summary of each of the individual papers with particular focus on the issue being explored in each paper, the general approach taken, and how the work fits into the overall thesis. Section 5.1 first describes how correlated cue sensitivity might be enhanced by probabilistic feedback and how this idea was tested in Chapter 6. Section 5.2 describes how increased correlational cue sensitivity in probabilistic environments might lead to heterogeneous knowledge (i.e., knowledge partitioning); this hypothesis was explored in Chapter 7. Section 5.3 provides the context for Chapter 8, which calls for a clarification in the terminology used in the literature regarding correlated cues in light of special role of selective attention. Section 5.4 briefly describes knowledge partitioning in the applied setting. Finally, Section 5.5 describes a particular assumption of the leading model of probabilistic category learning and how this assumption was tested in Chapter 10.

5.1 Correlated Cues and Probabilistic Feedback

This paper examines whether the hypothesized increased attention shifting that results from the unavoidable error in probabilistic environments results in increased sensitivity to non-diagnostic correlations. To explain, attention is predicted to shift amongst the stimulus dimensions whenever a response error is made (Kruschke & Johansen, 1999). Probabilistic feedback increases the level of error by preventing perfect categorization
performance; hence, probabilistic feedback should increase the level of attention shifting. If sensitivity to correlated cues is determined by whether or not the correlated cues are attended then probabilistic feedback should result in increased sensitivity to non-relevant cue correlations. By contrast, a deterministic feedback condition should show no sensitivity to non-diagnostic information.

5.1.1 Summary of Method

Participants were trained to categorize items comprised of four stimulus features (three filled or open circles and a fourth color dimension). A correlation was created by withholding items from presentation during a training phase such that the value of one of the circle dimensions was perfectly correlated with the color of the stimulus. Participants were provided with either perfect category feedback (i.e., the deterministic condition) or with imperfect category feedback (i.e., the probabilistic condition). The withheld items were reinstated during transfer tests and correlational sensitivity was assessed by a number of transfer tests which compared training items (which consistently maintained the correlation) with new transfer items (which broke the correlation).

5.1.2 Summary of Results

The results indicated that participants in the probabilistic condition demonstrated greater sensitivity to the correlated cues than the deterministic condition. Further analysis was conducted by fitting computational categorization models to the data from both conditions. The model analysis revealed that probabilistic feedback resulted in a spreading of attention across stimulus features as identified by the increased utilization of multiple stimulus components, increased ability to complete missing correlated features and diffuse attention weights in an exemplar model.
5.1.3 Summary of Discussion

This paper demonstrates that it is possible to observe correlational sensitivity in category learning tasks and implicates selective attention as the mechanism behind the acquisition of correlational knowledge. Interestingly, full sensitivity to correlated cues was not observed in these experiments. The probabilistic condition was only partly sensitive to the correlated cues. Full sensitivity to the correlation would have meant that a different categorization strategy would have been applied in each level of the non-relevant color cue, much like knowledge partitioning, which has been demonstrated in deterministic category learning and function learning. The next paper addresses whether knowledge partitioning can arise in probabilistic categorization.

5.2 Knowledge Partitioning in Probabilistic Category Learning

One reason that knowledge partitioning (i.e., full sensitivity to the correlated cues) might not have been observed in Chapter 6 is the use of binary (i.e., discrete) stimulus dimensions. Previous knowledge partitioning results have used a discrete context cue and continuous dimension such that each value of the context cue was associated with a different region of the stimulus space.

5.2.1 Summary of Method

To test knowledge partitioning, participants were trained to categorize stimuli comprised of two dimensions, an irrelevant discrete context cue (e.g., stimulus color) and a relevant continuously valued cue (e.g., levels of shading). In one context, the objective probability of category A linearly increased as shading increased; in the other context, the objective probability of category A linearly decreased as shading increased. Following training, participants were tested on all levels of shading in both contexts as well as items with shading levels outside of the training region. Knowledge partitioning was identified if a
5.2.2 Summary of Results

The results indicated that one group of participants ignored context and responded to the transfer items in a manner consistent with generalization based on exemplar similarity (confirmed by computational modeling). A second proportion of participants, however, extrapolated linearly along the objective probability function in a manner consistent with the use of two contrasting rule boundaries (i.e., in one context, participants split the items with probabilistic reinforcement to category A below .5 and above .5 with one rule boundary but they applied a different rule boundary in the other context). The use of multiple strategies to break a task down into smaller components is consistent with previous knowledge partitioning results. The existence of extrapolation outside of the learned training values is inconsistent with a selective attention account of learning. Instead, participants appeared to rely on context to gate the application of a rule boundary.

5.2.3 Summary of Discussion

Full correlational sensitivity in these experiments (i.e., knowledge partitioning) provides a boundary condition for the selective attention account correlational sensitivity. Interestingly, knowledge partitioning was correlated with a measure of fluid intelligence, such that, participants with higher mental capacity were less likely to demonstrate knowledge partitioning. This suggests that knowledge partitioning might be a way to strategically reduce the complexity of difficult tasks. This idea is explored further in Chapter 11.

5.3 Correlated Cues and Selective Attention

One of the explanations for the mixed evidence of correlational sensitivity in category learning is that multiple definitions of ‘correlation’ have been used to study correlational
sensitivity. These definitions include: a) statistical properties of bivariate normal distributions, b) relevant XOR cues, c) diagnostic within-category co-occurrences, d) prior knowledge consistent correlations, and e) the correlation of a discrete context cue with areas of the category space. This paper explores these different definitions and how selective attention might account for the varying results.

5.4 Applied knowledge partitioning

This paper explores how the conception of knowledge as multiple, encapsulated parcels, which are selected based on superficial, non-relevant context cues, can be utilized to inform behavior in applied settings. Evidence from studies of expertise, empirical demonstrations of knowledge partitioning and transfer of training results are used to support this view of knowledge.

5.5 Error Discounting

In addition to the increase in attention shifting which motivated Chapters 6 and 7, a corollary assumption of learning in probabilistic environments is that learning is attenuated over time as participants grow accustomed to an unavoidable level of error (Kruschke & Johansen, 1999). This assumption has been put forward as an explanation for the slowed learning effects in probabilistic categorization (i.e., when a relevant cue is introduced after a period of learning it is underutilized compared to a relevant cue of equivalent validity that is introduced at the outset). However, the attenuation rates reported in Chapter 7 were below the levels cited in Kruschke and Johansen (1999) and little empirical data exists to confirm the assumption of slowed learning. This paper explores whether error attenuation is a real phenomenon by training participants to categorize items receiving probabilistic feedback. The experiments test whether participants are able to track changes in the feedback contingencies that occur at various points during training. In all cases, participants are able to adjust their performance to the change but only after a significant
delay. Computational modeling confirmed that including a learning attenuation parameter significantly improved the model fits.
Chapter 6

Probabilistic feedback increases sensitivity to correlated cues.

*Paper 1 (Refereed)*

Probabilistic Feedback Increases Sensitivity to Correlated Cues in Categorization

Daniel R. Little and Stephan Lewandowsky

University of Western Australia

Daniel R. Little

Department of Psychological and Brain Sciences

Indiana University

1101 East Tenth St

Bloomington IN 47405-7007

e-mail: daniel.r.little@gmail.com

fax: +1 (812) 855 4691
**Abstract**

Despite the fact that categories are often composed of correlated features, the evidence that people detect and use these correlations during intentional category learning has been overwhelmingly negative to date. Nonetheless, people show evidence of correlational sensitivity during other categorization tasks (e.g., feature prediction; Chin-Parker & Ross, 2002). A conventional explanation holds that category learning tasks promote rule use which discards the correlated-feature information; whereas other types of category learning tasks promote exemplar storage which preserves correlated-feature information. Contrary to that common belief, we report two experiments that demonstrate that using probabilistic feedback in an intentional categorization task leads to a diffusion of attention across stimulus cues (Experiment 1) and can result in sensitivity to correlations among non-diagnostic cues (Experiment 2). Deterministic feedback eliminates correlational sensitivity by focusing attention on relevant cues. Computational modeling reveals that exemplar storage coupled with selective attention is necessary to explain this effect.

**Keywords:** Probabilistic Categorization, Correlated Cues, Selective Attention, Exemplars vs Rules
Probabilistic Feedback Increases Sensitivity to Correlated Cues in Categorization

The categories that we form about the world are undeniably comprised of correlated attributes. For example, members of the category birds have wings, can fly, have feathers, have beaks, and lay eggs. Notwithstanding obvious antipodean exceptions, the prototypical bird would tend to be comprised of all of these correlated “birdy” attributes (Murphy, 2002). Despite the consensus that natural categories are in part defined by their correlated attributes, the evidence for whether people detect and use these correlated features when learning new categories is at best mixed, with correlational knowledge being observable in some circumstances but not in many others.

To date, several different definitions of correlated cue have been used in the literature; ranging from the statistical correlation between continuous stimulus features (Anderson & Fincham, 1996; Crawford, Huttenlocher, & Hedges, 2006; Huttenlocher, Hedges, & Vevea, 2000; Thomas, 1998); to the interaction of stimulus features in an XOR relationship (Medin, Altom, Edelson, & Freko, 1982; Wattenmaker, 1991); to non-diagnostic cues marking sets of exemplars (e.g., subtyping; Bott & Murphy, 2007), regions of the stimulus space (e.g., knowledge partitioning; Little & Lewandowsky, in press; Yang & Lewandowsky, 2004), or values of another continuous cue (Minda & Ross, 2004); to diagnostic between-category feature correlations in which the correlation between features is predictive of the category outcome (Wattenmaker, 1993; Experiment 3); and to non-diagnostic within-category feature correlations in which the correlation provides information about the category members but does not determine category membership (Ahn, Marsh, Luhmann, & Lee, 2002; Chin-Parker & Ross, 2002; Malt & Smith, 1984). Here we are concerned with the final usage, namely, non-diagnostic within-category feature correlations.
Whether or not people have access to non-diagnostic feature correlations is important for at least two reasons: First, knowledge of the statistical properties of the environment facilitates prediction of future events. For example, knowing that “being small” is correlated with the property of “singing” permits the correct inference that a Willie Wagtail sings whereas an Emu does not, notwithstanding the fact that both animals can be classified as birds without knowledge of either their size or their vocalizations. Second, the use or non-use of correlational knowledge provides a particularly strong test for models of category learning because some classes of models predict that correlational knowledge should be accessible whereas others do not. For instance, a rule or prototype supplies knowledge of the category without providing information about the relationship between features (e.g., that small birds are more likely to sing than large birds). By contrast, an exemplar model potentially provides access to this information because all features are stored for each instance: in consequence, information about bird size and song could be retrieved later and compared to produce the correct inference.

To date, a clear picture of correlated-cue knowledge has failed to emerge, owing in part to the use of widely divergent methods. Relevant research has been conducted in two broad paradigms; namely, category learning tasks in which the participant intentionally seeks to learn experimenter-defined categories via error-correcting feedback, and category usage tasks in which the participant must use category knowledge to achieve some other goal. The basic finding is that correlated-cue knowledge can be identified in the latter type of task but not the former (Ahn et al., 2002; Chin-Parker & Ross, 2002; McNorgan, Kotack, Meehan, & McRae, 2007; Wattenmaker, 1993). To foreshadow the remainder of this article, we first review these two types of tasks and highlight differences in selective attention as a possible causal factor that determines whether or not correlated cues are detected. On this assumption, any manipulation which affects selective attention should also affect knowledge of correlated cues; we explore this argument in two experiments.
which respectively demonstrate that a) probabilistic feedback results in the utilization of more stimulus features than deterministic feedback, suggesting that under probabilistic feedback conditions participants spread their attention; and b) that probabilistic feedback leads to sensitivity to non-diagnostic correlated cues. Two models that differ in their underlying representations (i.e., exemplar vs. rule) are then fit to the experimental data. The modeling reveals that correlational sensitivity is the result of a broadening of attention across multiple stimulus features that occurs in response to probabilistic feedback. We find that an exemplar representation is required to capture correlational sensitivity when present.

Category Usage vs. Category Learning

Category usage tasks require either the application of category knowledge or the processing of category exemplars without explicit instructions to learn the categories’ criterial attributes. The category-usage paradigm encompasses an array of tasks such as feature prediction, incidental learning, and typicality ratings. Generally, all of these tasks reveal sensitivity to correlated attributes. For example, when participants rate the typicality of feature pairs for a particular category, typicality tends to be judged higher if the features are causally connected by prior knowledge (Ahn et al., 2002; Malt & Smith, 1984; McNorgan et al., 2007). Thus, the bird features “has a large beak” and “eats fish” are judged to be more typical than “flies” and “chirps” (Ahn et al., 2002), presumably because participants can explicitly theorize about the correlation (i.e., “a large beak” allows the bird to “eat fish”). Features which are not causally dependent are rated to be less typical (e.g., “flying” and “chirping”).

People also become sensitive to correlational information when the learning tasks require processing of all of the stimulus features, such as during feature prediction (Anderson & Fincham, 1996; Chin-Parker & Ross, 2002). Chin-Parker and Ross (2002)
trained participants either in a category learning task, which required participants to respond with the missing category label, or in a feature inference task, which required participants to respond with a missing stimulus feature. The only difference between the tasks was the nature of the missing information that had to be inferred. Chin-Parker and Ross (2002) found that participants trained in the feature-inference task showed greater correlational sensitivity across four different tests than participants trained in the category learning task. In the category learning tasks, people failed to detect the correlated cues, and this failure turns out to be a pervasive attribute of category learning.

For example, Wattenmaker (1991) instructed participants to learn as much as possible about members of a single category whose four stimulus features involved several non-diagnostic correlations. After training, participants demonstrated little ability to access information about these correlated cues and were aware only of the diagnostic features and configurations. Similar results have been reported by Murphy and Wisniewski (1989), who found that the effects of prior knowledge far outweigh experimental training of family-resemblance categories with co-occurring features.

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

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

Selective Attention and Correlational Sensitivity

Most theories of attention in category learning assume that people attend to those features that will maximize accuracy (Kruschke, 1992, 2001, 2003; Kruschke & Johansen, 1999; Lamberts, 1995; Nosofsky, 1986). This view is supported by multiple streams of evidence, including the relative ease of learning when fewer cues are diagnostic (Kruschke, 1993; Shepard, Hovland, & Jenkins, 1961); blocking and highlighting (Kruschke, 2003; Kruschke & Blair, 2000); the fact that intra-dimensional validity shifts (i.e., when the response validities of a cue change but its identity does not) are easier to learn than extra-dimensional shifts (i.e., when a different cue becomes valid; Kruschke, 1996); and eye-gaze tracking experiments that show that eye-gaze direction is correlated with feature validity (Kruschke, Kappenman, & Hetrick, 2005; Rehder & Hoffman, 2005b, 2005a).

In addition, one theory of attention-shifting proposes that when participants make a classification error, their attention is briefly and fleetingly shifted from the learned diagnostic features to other features of the current stimulus (Kruschke, 2001; Kruschke &
Johansen, 1999). By implication, attention shifting may be particularly prevalent in categorization tasks with probabilistic feedback, in which an item’s category membership cannot be predicted from trial to trial with absolute certainty; only long-term trends can be learned (e.g., “this item belongs to category A 70% of the time”). In consequence, perfect performance cannot be achieved and participants are forced to accept some level of error throughout the task.

The persistence of error, and the attention-shifting it entails, may explain why participants typically under-utilize valid cues in probabilistic categorization and instead over-utilize irrelevant cues (see, e.g., Castellan Jr., 1973; Edgell et al., 1996; Edgell, 1980; Kruschke & Johansen, 1999). By implication, if attention is required to detect feature correlations, the use of probabilistic feedback, which is known to increase the prevalence of rapid attention shifts, might in turn enhance correlational sensitivity. The experiments that follow pursue two goals with respect to this hypothesis. The first experiment offers an existence proof of key differences between probabilistic and deterministic feedback conditions when people are given a choice of two valid cue components. The second experiment demonstrates that probabilistic feedback engenders sensitivity to a non-diagnostic correlation in an intentional category learning task.

Categorization tasks are often characterized by large individual differences in the strategies used to solve the task (see e.g., Erickson & Kruschke, 1998; Little & Lewandowsky, in press; Rouder & Ratcliff, 2004; Thomas, 1998; Yang & Lewandowsky, 2003, 2004). This is particularly relevant in the present case because prior research examining correlational sensitivity has relied on complex stimulus spaces with multiple relevant and non-relevant correlated cues (see e.g., Chin-Parker & Ross, 2002; Wattenmaker, 1991, 1993). It follows that correlational sensitivity may be a by-product of other types of responding—for example, a non-diagnostic correlation may be a constituent part of an over-arching rule involving all available features. Hence, the following
Correlated Cues in Probabilistic Categorization

experiments involve a simpler design to permit accurate identification of the cues that drive people’s decisions in addition to detecting correlational sensitivity.

**BEHAVIORAL EXPERIMENTS**

**Experiment 1**

The principal aim of Experiment 1 was to examine whether people would spread their attention over multiple cues when categories were probabilistically reinforced. The experiment compared performance with deterministic and probabilistic feedback on a categorization problem involving four features. The problem involved two compound cues (i.e., pairs of features) that were equally valid but differed in their intra-pair relational similarity. Relational similarity refers to the fact that “O!O” is considered more similar to the target “XMX” than “XXA”, despite the fact that “XXA”—but not “O!O”—shares features with the target (Goldstone, Medin, & Gentner, 1991). There is evidence that people favor cues that are relationally similar in categorization tasks (e.g., Little & Lewandowsky, in press); it follows that in this experiment people may also preferentially use the relationally similar compound cue.

Participants were trained to categorize stimuli comprised of four binary features. Three of the features (called X, Y, and Z) were instantiated by circles that were either open or filled (hereafter called shading). The fourth feature (called C) was instantiated by the color common to all circles (red or green). Table 1 summarizes all stimuli and the category structure; Figure 1 shows several examples of the stimuli. The table identifies two diagnostic XOR cues: one comprised of features of the same type (i.e., two circles varying in shading; Y & Z; high relational similarity) and another which was equally diagnostic but contained features of different types (i.e., color and shading; Y & C; low relational similarity). The conjunction of either of these paired cues predicted the category outcome either perfectly (deterministic condition) or probabilistically (probabilistic condition).
During analysis, focus was on examination of people’s distribution of attention across features. Attention profiles were investigated by considering people’s responses on a categorization transfer test and a similarity rating task.

Method

Participants.

Forty-seven University of Western Australia students received partial course credit or $10 remuneration for participation. Participants were randomly assigned to the deterministic condition (N = 21) and the probabilistic condition (N = 26).

Stimuli and apparatus.

The stimuli were composed of three binary-valued dimensions, X, Y, and Z, instantiated by shading of circles (i.e., open or filled), and a fourth binary-valued dimension, instantiated by the color of all circles (C), which was perfectly correlated with the Z dimension during training (see Figure 1). The assignment of physical stimulus features to the abstract X, Y, Z dimensions was counterbalanced across participants. The values of each feature assignment were randomized, so that, for example, a filled circle might represent a value of 0 and an open circle might represent 1, or vice versa. The color of the stimulus was always mapped to the abstract C dimension, though the value of the C was randomized for each participant (e.g., red might represent a value of 0 and green might represent 1 or vice versa). Depending on condition, the feedback during category learning was either deterministic or probabilistic (see Table 1 for a description of all stimuli).

In both conditions, the base-rate validity, or overall mean probability that a stimulus belonged to category A, \( P(A) \), was 0.5 (see Table 2). Category B probabilities were \( 1 - P(A) \). The component validities are expressed as deviations from this base validity. (see Kruschke & Johansen, 1999, for a detailed review of this calculation). For example, in
both conditions, averaging the probability of category A when dimension X had a value of 0 yields .50 (see Table 2); this average deviated from the base rate by .00 indicating that dimension X was not a valid predictor (hence, the validity of X was 0). The only stimulus components with any validity were the YZ and YC stimulus components. To illustrate, if features Y and Z are recoded as -1 and +1 (from 0 and 1, respectively) and multiplied, the average probability of category A when YZ equals 1 would return a value of 1.00 in the deterministic condition and a value of .75 in the probabilistic condition, yielding deviations from base rate of .50 and .25, respectively. Thus, the YZ component had a validity of .50 in the deterministic condition and .25 in the probabilistic condition. This computation of validity linearly decomposes the feedback proportions into orthogonal components. The validity of each feature then is the conditional probability of Category A given that the feature takes on a value of 1 minus the category base rate. Note that the probabilistic feedback reduces the validity of the valid cues. Because only stimuli that preserved the ZC correlation were shown during training, the categorization problem was isomorphic to a Shepard et al. (1961) Type II category space in both conditions.

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

*Design and procedure.*

Following precedent (Kruschke & Johansen, 1999), detailed instructions outlining the probabilistic nature of the training task were given to participants in the probabilistic condition at the outset.

In both conditions, training consisted of 8 blocks of trials, each involving 5 presentations of the 8 training items in a different random order for a total of 320 trials.
On each training trial, participants had to categorize the stimulus into one of two categories, “F” or “J”, by pressing the corresponding keys on the keyboard. (Each stimulus was ostensibly presented as a strand of Alien DNA to be grouped into two different species of alien). After a response, feedback (“CORRECT” or “WRONG”) was presented underneath the stimulus. Feedback was generated randomly from the probabilities shown in Table 1 and remained visible for at least 1 s, after which participants could press the spacebar to advance to the next trial. If participants did not press the spacebar, feedback remained visible for 15 s. In addition to feedback for each trial, at the end of each block the computer also displayed the percentage correct for the previous block.

Transfer tests contained two presentations of the 16 test items in random order with a 500 ms blank interval between each response and the next item. Transfer trials were identical to training trials except that feedback was withheld and all of the items shown in Table 1 were presented. Instructions for the transfer test were presented following training.

Following the category transfer test, each participant completed a similarity rating task by providing ratings for all 120 pairwise combinations of the 16 transfer stimuli. Each stimulus pair was presented simultaneously and remained on screen until a response was entered. Responses were made using an 8-point scale (i.e., 0 = Least Similar to 7 = Most Similar).

Results

Training performance.

To ensure that only motivated and able participants were included in the analysis, we adopted the criterion of Edgell et al. (1996) and screened the data from both experiments for long strings of consecutive identical responses, using a cutoff of 30 identical responses. No participants exceeded this cutoff; hence, all participants were
Correlated Cues in Probabilistic Categorization 13

retained for analysis.

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

\[ PM_{ij} = \left( \frac{1}{N} \sum_{j=1}^{N} [P(A|j) - R_i(A|j)] SI_j \right) + ADJ_c, \]  

where \( P(A|j) \) is the probability of receiving feedback for category A given item \( j \) (see Table 1), \( R_i(A|j) \) is participant \( i \)'s proportion of 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). These probability matching (PM) scores were averaged across all items for each participant. The scores were then adjusted, \( ADJ_c \), by adding +1 to the deterministic condition and +.75 to the probabilistic condition; this adjustment means that for both conditions, a PM score of .75 indicates that the response proportions were identical to the objective probabilities of the probabilistic condition. A PM score greater than .75 indicates overshooting of the training probabilities with \( PM = 1.00 \) indicating perfect maximizing. For both conditions, a score less than .75 indicates undershooting of the training probabilities (with \( PM = .50 \) indicating chance performance and \( PM < .50 \) indicating a reversal of the training probabilities).\(^3\)

The PM scores of both conditions are shown in Figure 2 (see panel A). Responding in the deterministic condition quickly reached accurate levels. The probabilistic condition
approached the objective probabilities although there was a remaining tendency to undershoot at the end of training. The clear upward trend indicates that both conditions demonstrated learning across training blocks. In confirmation, paired-sample $t$-tests revealed significantly higher PM scores in the eighth and final block than in the first block for the deterministic condition, $t(20) = 10.25, p < .05$, Cohen’s $d = 2.24$, and the probabilistic condition, $t(25) = 3.27, p < .05$, Cohen’s $d = 0.64$.

Figure 3 (panels A and B) shows the aggregate training performance from the final two training blocks of both conditions compared to the feedback probabilities for each training stimulus. The figure confirms that responses in the deterministic condition matched the feedback probabilities more closely than in the probabilistic condition.

**Transfer performance.**

Transfer performance was statistically examined by computing the utilization of each stimulus component for each participant. Utilization is computed in the same way as validity (i.e., deviation from base rate validity), with the exception that response proportions are used in the utilization calculation whereas the feedback probabilities are used in the validity calculation. Figure 4 shows the utilization rates of each of the stimulus components for the two conditions.

In both conditions, the utilizations of the two valid component, YZ and YC, deviated notably from zero. Two additional points are worth noting: First, the probabilistic condition utilized the YZ component much less than the deterministic condition; an independent-samples $t$-test corrected for unequal sample size and unequal variance confirmed this difference, $t(40.99) = 3.87, p < .01$, Hedges’s $g = 0.33$. Second, the differences between the utilizations of YZ and YC were much larger in the deterministic condition, Cohen’s $d = 1.75$, than in the probabilistic condition, Cohen’s $d = 0.26$. The difference in these effect sizes, Cohen’s $q = 0.66$, is of a moderate size (Cohen, 1988). Overall, the deterministic condition clearly preferred YZ over YC. This trend was also
followed in the probabilistic condition, but the difference was much smaller.

**Similarity ratings.**

We computed the difference in similarity ratings between items that differed on only one feature (e.g., two stimuli whose X, Y, and Z dimensions were identical but which differed on C). Those pairwise differences were aggregated across all possible stimuli for each feature (e.g., across all combinations of X, Y, and Z values). Any feature that was an important determinant of the similarity ratings would be expected to yield a large difference between stimuli that differed on this one feature only. The use of similarity ratings to infer people’s representations after category learning has ample precedent (e.g., Goldstone, Lippa, & Shiffrin, 2001; Livingston, Andrews, & Harnad, 1998).

The effect of higher-level components (e.g., the conjunction of Y and Z) were computed in an analogous manner to utilization and validity; namely, by contrasting similarity ratings between items that would be classified into one or the other category on the basis of only that component and regardless of what was seen at training. For instance, the XY component on its own predicts that the items CN1, CN4, CN5, and CN8 belong in one category whereas items CN2, CN3, CN6, and CN7 belong in the other category; hence, the role of that component was evaluated by comparing the average within-category pairwise similarities (e.g., CN1 vs. CN4 or CN2 vs. CN3) with the average between-category pairwise similarities (e.g., CN1 vs. CN2 or CN3 vs. CN8).

The average differences in similarity ratings for all components are shown in Figure 5. The figure shows that the deterministic group was largely influenced by the diagnostic YZ component. This tendency also arose in the probabilistic condition but to a considerably lesser degree. The deterministic condition was also influenced by several higher-level components that contained the YZ component (i.e., XZC, YZC, and XYZC; see Figure 5). By contrast, the probabilistic group was affected by several single features (i.e., X, Y, and Z; the C feature was marginally different from zero,
\( t(25) = -2.03, p = 0.05 \) and several higher-level components (i.e., XY, XC, YZ, YC, ZC, XYC, YZC, and XYZC), indicating that a larger number of components influenced the similarity ratings in this condition. Overall, 13 out of the 15 components showed significant differences in the probabilistic condition compared to only 5 in the deterministic condition. (Using a Bonferroni correction for multiple comparisons, this drops to 1 component for the deterministic condition and 7 for the probabilistic condition). This result confirms that people’s attention in the probabilistic condition was distributed more diffusely.

Discussion

The principal goal of Experiment 1 was to examine whether probabilistic reinforcement leads to a more diffuse allocation of attention across multiple features than deterministic categorization under otherwise identical circumstances. This goal was met. In the deterministic condition, people used only one of the diagnostic components (YZ) and ignored other possible cues or their component (including the alternative diagnostic component YC). In the probabilistic condition, by contrast, people exhibited a much wider attention profile. Consistent with previous work comparing cue utilization of differentially valid cues (cf., Edgell, 1978, 1980; Edgell et al., 1996; Kruschke & Johansen, 1999), people in the probabilistic condition showed decreased utilization of the valid YZ component compared to the deterministic condition. The reduced reliance on YZ was accompanied by an increased utilization of the equally diagnostic YC component, indicating that the introduction of probabilistic feedback led to greater diffusion of attention. The utilization data were mirrored by the similarity ratings, which also showed that participants in the probabilistic condition attended to a wider array of stimulus components than people in the deterministic condition.

In addition to showing that probabilistic reinforcement led to a broader attention
profile, Experiment 1 also revealed which of several equally-valid components people will prefer when their attention profile is not broad but instead narrowly focused by deterministic reinforcement. Participants in the deterministic condition overwhelmingly utilized the component involving features of the same type (i.e., two circles; YZ) rather than using a formally identical rule involving features of two different types (i.e., one circle and color; YC). Thus, when given the choice and when attention is not diffused by probabilistic reinforcement, people use relational similarity to choose between two valid components.

A potential by-product of a broad attention profile is that people become sensitive to non-diagnostic correlations among cues. To examine this issue, Experiment 2 again compared deterministic and probabilistic categorization but using a category structure that included a correlation among non-diagnostic cues. To identify any correlational sensitivity that might arise, Experiment 2 included a feature-prediction test in addition to the conventional categorization transfer test.

**Experiment 2**

Experiment 2 used a stimulus space that contained a single diagnostic component (XY) and a single non-diagnostic correlated cue (zc). The constituent dimensions (Z and C) of the non-diagnostic correlation (zc) did not enter into the diagnostic component (i.e., XY; see Table 2). However, the zc correlation instead formed part of a diagnostic over-arching XYZC component because it is logically impossible to create a non-diagnostic correlation among two cues in a four-dimensional stimulus space without contributing to another diagnostic component, such as the over-arching diagnostic correlation among all cues as in the present study.

By implication, participants could learn the structure in Experiment 2 by relying on one or the other valid component: People could either form an XOR rule involving X and
Y or they could form a rule involving XYZC. The latter is rather complex; the rule would note that any item whose color is “0” (see Table 1; \( C = 0 \) may refer to green or red depending on counterbalancing) with 0 or 2 circles filled and any item of color “1” with one circle filled belongs to category A; otherwise, the item belongs to category B.

Utilization of XYZC subsumes sensitivity to the \( zc \) correlation: Hence, to establish whether people are sensitive to the non-diagnostic correlation \textit{per se} or whether they exhibit that sensitivity consequent upon use of the over-arching diagnostic XYZC component, Experiment 2 involved a partial-stimulus completion test in addition to the conventional transfer test.

The transfer test was identical to that used in Experiment 1; that is, participants were shown both consistent and inconsistent items and were required to make a category response. Utilization rates on this test can index which of the two valid cues (and potentially also non-diagnostic cues) participants use. Any knowledge of the \( zc \) correlation would be reflected in significant utilization of \( zc \).

The feature completion test involved presentation of each training item in black along with a category label. Participants were required to complete the stimulus with the appropriate color. If participants have knowledge of the non-diagnostic \( zc \) correlation, then they should consistently respond with the color that was present throughout training. Conversely, if participants are using the diagnostic XY component without knowledge of \( zc \), or if participants are using the XYZC component, then the completion responses (when averaged across all items) should match the trained colors only 50% of the time: If people rely on XY without knowledge of the \( zc \) correlation, they can only guess the color in which a stimulus was presented, thus leading to chance performance. If people rely on the XYZC component, their performance should nonetheless be at 50% because the XYZC-rule makes contrasting predictions depending on the category that is presented along with the partial item. For example, if the item “100?” is presented labeled
as Category B, the XYZC component mandates that the color be “0”; see item CN2 in Table 1. However, the same item labeled as Category A should lead to the opposite response (see Table 1 and the verbally stated rule above). It follows that any above-chance responding with the trained color on the feature completion test is uniquely and unambiguously indicative that people have knowledge of the non-diagnostic zc correlation independent of any involvement of a valid component.

Method

Participants and apparatus.

Sixty-eight University of Western Australia students received partial course credit or $10 remuneration for participation and were randomly assigned to the probabilistic condition (N = 38) or the deterministic condition (N = 31). The surface features of the stimuli and apparatus were identical to Experiment 1.

Design and procedure.

The stimuli and category structure are summarized in Table 1. The training phase proceeded in an identical manner to Experiment 1; however, there were two additional training blocks of 5 item repetitions per block, yielding a total of 400 training trials. Following the training phase, participants then completed the feature completion task followed by the transfer categorization test. Following these tests, participants then completed a typicality rating test; however, the results of this test added little to the analysis of the other tests and we elected not to report these results for the sake of brevity. The similarity rating task used in Experiment 1 was omitted here because it could shed little additional light on the effect of greatest interest, namely sensitivity to the non-diagnostic correlation. Test instructions were presented after the learning phase ahead of each test.

For the feature completion task, participants were shown the stimuli listed in
Table 1 in black along with a category label and had to respond by providing the missing color (C). For each response, the participant first selected the color; this action allowed the participant to preview the item in the chosen color prior to confirming that selection (or clearing it to choose an alternative response). After the response was confirmed, there was a 500 ms blank interval before the start of the next trial. Each of the eight possible combinations of X, Y, and Z were shown 8 times with each category label, yielding 128 feature completion trials. The stimulus presentations were randomized anew for each participant.

The transfer categorization test was identical to the transfer test in Experiment 1; however, in this study there were 10 repetitions of the 8 consistent and 8 inconsistent test items in random order for each participant (see Table 1).

Results and Discussion

Training performance.

Three participants in the probabilistic condition who had one or more consecutive strings of more than 30 identical responses were removed from the analysis, leaving 35 participants in the probabilistic condition and 31 participants in the deterministic condition.

Training performance is shown in panel B of Figure 2. As in Experiment 1, learning in the probabilistic condition proceeded more slowly than in the deterministic condition. Nonetheless, the average responses in the final three training blocks were accurately matched to the feedback provided at training, indicating that both conditions ultimately learned the task (see Figure 6). In particular, the additional training in this study enabled participants in the probabilistic condition to match the training probabilities more closely than in Experiment 1.

Transfer performance.
Transfer performance was again examined by computing utilizations of each stimulus component (see Figure 7). In both conditions, we again found that both of the valid components (XY and XYZC) were significantly utilized (significance can be assessed from the figure; any confidence interval that does not span 0 is equivalent to a significant \( t \) with \( \alpha = .05 \). The deterministic condition additionally utilized the Z dimension (Cohen’s \( d = 0.50 \)); though its extent of utilization was considerably less than the utilization of XY (Cohen’s \( d = 1.10 \)) and XYZC (Cohen’s \( d = 0.73 \); see Figure 7).

The probabilistic condition likewise utilized the valid XY (Cohen’s \( d = 0.72 \)) and XYZC components (Cohen’s \( d = 0.67 \)). Unlike in the deterministic condition, probabilistic reinforcement also led to utilization of the X dimension (Cohen’s \( d = 0.44 \)), the XC component (Cohen’s \( d = 0.45 \)), the YC component (Cohen’s \( d = 0.39 \)), and the non-diagnostic \( zc \) component (Cohen’s \( d = 0.36 \)), again demonstrating the spread of activation across stimulus components that arises with probabilistic reinforcement. Note that the effect sizes were smaller for the non-diagnostic components (X and \( zc \)) than for utilizations of the valid components.

*Feature completion data.*

Each participant contributed to 128 feature completion trials, yielding a chance cutoff of 73 (\( \alpha = .05 \), using a binomial distribution with \( N = 128 \) and \( P = Q = .5 \)). Thus, any participant who completed the missing feature as it appeared during training on more than 73 trials was considered to be above chance and hence exhibited knowledge of the non-diagnostic correlation. Reliance on either of the valid components, XY or XYZC, does not permit above-chance performance on this test.

There were 3 participants (out of 31) in the deterministic condition whose scores were above chance, compared to 17 such participants (out of 35) in the probabilistic condition. The individual results were mirrored at the aggregate level, with average performance in the deterministic condition being indistinguishable from chance.
(M = .53, SD = .09), t(30) = 1.95, p > .05, in contrast to the probabilistic condition in which performance was significantly above chance (M = .62, SD = .15),
t(34) = 4.62, p < .05, Cohen’s d = .79. A one-tailed t-test comparing the two groups showed that the probabilistic group had significantly higher feature completion scores than the deterministic group, t(64) = 2.83, p < .01, Cohen’s d = .71. Figure 8 captures the differences between conditions graphically, by showing the distributions of feature completion scores in both conditions. It is clear from the figure that whereas the vast majority of deterministic participants clusters at or right below .5, in the probabilistic condition half of the distribution lies clearly above .5 (the fact that the other half of distribution in the probabilistic condition fails to exceed .5 is taken up in the modeling section; for now, the data are sufficiently clearcut to permit conclusions without consideration of individual variations.)

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

This finding is particularly striking in light of the results of Experiment 1; in that experiment, when presented with two components of equal validity, participants preferentially used the pair of cues that had greater relational similarity. This preference for relational similarity did not prevent participants in the present study from learning a non-relevant correlation (\( zc \)) comprised of features of different types.

What, then, explains the sensitivity to non-diagnostic correlations observed with probabilistic reinforcement? What representations are required to account for the results
of Experiment 1? Previous research has suggested that knowledge of correlated cues is the outcome of memory for individual exemplars (Wattenmaker, 1991, 1993). This interpretation is also attractive in the present case because any participant who memorized the stimuli would presumably be able to recall those exemplars later for the feature completion task. In order to explore the idea that exemplar memory underlies correlational knowledge, we now compare how an exemplar model and two rule models handle the key result of Experiment 2; namely, the correlational sensitivity in the probabilistic (but not the deterministic) condition.

**COMPUTATIONAL MODELING**

*Overview*

We opted to contrast an exemplar model, the generalized context model (GCM; Nosofsky, 1986; Nosofsky & Johansen, 2000), with two rule models based on the set-of-rules model (SRM) introduced by Johansen and Kruschke (2005). Our choice was informed by previous work which has variously pointed to the centrality of exemplar storage in any demonstration of correlational sensitivity (Wattenmaker, 1993, 1991) and to the importance of rule-based representations in feature-prediction tasks (Johansen & Kruschke, 2005).

The SRM’s mechanisms overlap considerably with those of the GCM and it differs primarily with respect to what is stored in memory. Whereas the GCM stores all previously encountered exemplars, the SRM stores only the relevant feature-to-category associations. The use of two closely related but conceptually quite distinct models has two advantages: First, it allows exploration of our main theoretical questions—is selective attention important in determining correlational sensitivity? Can correlational sensitivity be modeled without reliance on the same correlation for categorization?—without commitment to a specific underlying representation. Second, comparing rule-based to
exemplar-based variants addresses the assumption that exemplars are central to cor relational sensitivity.

We next outline the two models and describe how they were fit to the categorization and feature prediction transfer data of Experiment 2. To foreshadow our conclusions, the categorization and feature prediction data were characterized by large individual differences. However, regardless of these differences, the deterministic condition turns out to be better characterized by a rule-based model whereas the probabilistic condition turns out to be better captured by an exemplar model. In addition to exemplar representations, the probabilistic condition requires a much broader attention profile (i.e., attention to more than just the relevant cues) than the deterministic condition. Furthermore, representation of non-diagnostic correlated cues turns out to be necessary in the probabilistic condition but not in the deterministic condition.

**GCM**

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

In the GCM, a stimulus activates all previously encountered stimuli stored in memory according to:

\[ s_{ij} = \exp(-cd_{ij}), \]  

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

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

where \( x_{ik} \) is the value of dimension \( k \) for test item \( i \) and \( x_{jk} \) is the value of dimension \( k \) for the stored exemplar \( j \), \( w_k \) is the attention weight for dimension \( k \), \( r \) indicates the distance metric, and \( P \) determines the form of the generalization gradient (\( P = 1 \), exponential or \( P = 2 \), Gaussian; Shepard, 1987). For all simulations, \( P \) was set to unity; however, \( r \) was allowed to vary between 0 and 2. Typically, \( r \) is set equal to 1 or 2 for modeling distances between separable (city-block distance) or integral (Euclidian distances) stimuli, respectively (Shepard, 1991). Values of \( r \) less than 1 indicate that the stimulus dimensions compete for attention, such that any attention to one dimension results in a greater loss of attention to other dimensions (Lee, 2008; Nosofsky, 1992; Shepard, 1987, 1991; Tversky & Gati, 1982). The attention weights, \( w_k \), are constrained to sum to one.

To apply the GCM to both categorization decisions and feature predictions, the category label was also instantiated as a feature dimension and given a value of 0 for category A and 1 for category B. Hence, the category similarity was computed across this dimension for the feature prediction task. Because the category label was the to-be-predicted feature, \( F_p \) in the categorization transfer test, it was excluded from the similarity comparison for that test. Conversely, the to-be-predicted feature, \( C \), was excluded from the similarity comparison in the feature prediction test.

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

\[ P(F_p = 0|i) = \frac{\left( \sum_{j \in F_p=0} s_{ij} \right)^\gamma}{\left( \sum_{j \in F_p=0} s_{ij} \right)^\gamma + \left( \sum_{j \in F_p=1} s_{ij} \right)^\gamma}. \]  

(4)
where the response scaling parameter, $\gamma$, allows responding to vary between probability-matching when $\gamma \approx 1$ and maximizing when $\gamma \gg 1$ (Ashby & Maddox, 1993; Nosofsky & Johansen, 2000).

**Rule Models**

The SRM is identical to the GCM in many respects but assumes that only feature-to-label associations are stored. Here we implemented two rule models motivated by the utilization data of Experiment 2; an XOR rule-model and an XYZC rule-model (hereafter, called the Rule-XOR and Rule-XYZC model, respectively). Importantly, like the GCM, these models estimate dimensional attention weights from the data.

Because the only valid stimulus components in Experiment 2 were the XOR and XYZC compounds, the feature-to-category mappings were also composed of higher-level compounds. For the Rule-XOR model, the stored associations are shown in Table 3. These stored associations replace the stored exemplars, $j$, in Equation 3. When a stimulus is presented for classification, all of the stored associations enter into the similarity calculation (see Equation 2). Only the features which are represented and are not being predicted (e.g., the category label for the categorization task and $C$ for the feature prediction task) are used to compute the distance between the stimulus and the stored associations (see Equation 3). These similarities are then converted to a response probability using Equation 4.

Strictly speaking, the associations displayed in Table 3 apply only to the deterministic condition. To fit the probabilistic condition, we assumed that only the most frequent rule-to-category association is stored, thus rendering Table 3 applicable also for the probabilistic condition. As shown in Table 3, the Rule-XOR model explicitly stores the XOR mapping from XY to categories A and B; hence, performance during the classification transfer test is
governed solely by the application of the XOR rule on X and Y. Accordingly, this model does not expect participants to be able to correctly predict the missing C feature during the feature completion task as neither Z nor C are explicitly represented in the model.

For the Rule-XYZC model, the stored associations are more elaborate than in either the Rule-XOR model or the GCM. Because the XYZC rule takes into account all four of the stimulus features, a full set of stored associations along with a full set of generalizations (i.e., all possible combinations of the stimulus features including those not shown during training) are represented (see Table 4). Note that the table entries do not represent exemplars, as in the GCM, but associations between each unique configuration of features and the associated response. In consequence, the table also contains entries for the novel transfer stimuli which were never shown during training but for which responses are prescribed by the XYZC rule. Commensurate with the demonstrable utilization of XYZC in Experiment 2, this model predicts that performance relies on utilization of dimensions X, Y, Z, and C for the categorization transfer test and X, Y, Z, and the Category label for the feature prediction test.

Parameter Estimation and Model Fitting

Each of the models had 5 attention weight parameters (i.e., $w_X$, $w_Y$, $w_Z$, $w_C$, and $w_{Cat}$ for the category label, which was necessary for the feature prediction task), a response scaling parameter, $\gamma$, and a distance metric parameter, $r$. The specificity parameter, $c$, was set equal to one as all of the attention weights were free to vary (cf. Johansen & Kruschke, 2005). Because dimensions which are not represented in the model do not enter into the similarity comparison, the $w_Z$ and $w_C$ parameters were effectively zero for the Rule-XOR model. Parameters were estimated for each participant separately, by fitting the models simultaneously to an individual’s categorization and feature-prediction responses. The best-fitting parameters were determined by maximizing
the binomial log-likelihood as follows:

\[ \ln L = \sum_i d_i \ln(p_i) + (n_i - d_i) \ln(1 - p_i) \]  \hspace{1cm} (5)

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

To account for the different number of parameters between the three contenders, we computed the Bayesian Information Criterion (BIC; Myung & Pitt, 2004) which adds a penalty term to the log-likelihood based on the number of free parameters and the size of the sample being fit:

\[ \text{BIC} = -2 \ln L + k \ln(n) \]  \hspace{1cm} (6)

where \( k \) is the number of free parameters and \( n \) is the number of cells that enter into the computation of the \( \ln L \).

**Strategy Analysis and Model Comparison**

To ensure that we were capturing only participants who consistently relied on one or the other compound cue rather than idiosyncratic guessing patterns, any participant with a PM score < .60 (indicating chance performance at the end of training using a lenient alpha level of .1 in the binomial probabilities with \( P = .5 \) for each training item) was excluded from the modeling; this left 29 participants from the deterministic condition and 27 participants from the probabilistic condition. The individual model fits for the two conditions are shown in Tables 5 and 6, respectively. The tables reveal clear differences in how the models apply across participants; for example, in the probabilistic condition, the GCM clearly provides the best account of participant 19, whereas the Rule-XYZC model
best accommodates the responses of participant 13 (the table entries are \(-\ln L\) values; hence smaller values indicate better fit). Following precedent (e.g., Little & Lewandowsky, in press; Yang & Lewandowsky, 2004), we sought to characterize these individual difference by reconsidering the data to identify different patterns of responding in the feature prediction task.

We conducted a \(k\)-means cluster analysis (with \(k = 2\)) on the average feature prediction score for each item. The starting points of the cluster analysis were selected randomly; the final cluster centroids returned by the \(k\)-means analysis were then compared to two ideal centroids based on (a) use of the XYZC cue and (b) knowledge of the \(zc\) correlation. As detailed in connection with Experiment 2, both XYZC and knowledge of \(zc\) unambiguously predict a specific response to each stimulus, thus creating two binary ideal response profiles that were compared to the recovered cluster centroids by correlation. Note that there is no a priori reason for the centroids recovered from the cluster analysis to necessarily resemble any of these ideal centroids.

The pattern of correlations is shown in Table 7 and permits two notable conclusions: First, the ideal XYZC centroid was correlated with a centroid from both the deterministic and probabilistic condition, which in turn were correlated with each other. This confirms the results of the utilization analysis, which pointed to the use of the XYZC compound in both conditions. The cluster analysis additionally identified the proportion of participants (approximately 40% - 50% in each condition) who primarily relied on the XYZC compound. Second, and more important, only the probabilistic condition returned a centroid which was correlated with the ideal profile associated with knowledge of the non-diagnostic \(zc\) correlation (roughly half of the people in that condition were best described by that centroid). The deterministic condition, by contrast, did not return a centroid aligned with knowledge of the non-diagnostic correlation.

To explore these differences and following relevant precedent (see e.g., Juslin, Jones,
we aggregated the data and model fits across participants within the groups identified by the \( k \)-means analysis. The aggregated model fits and parameters are presented in Tables 8 and 9, respectively. Not unexpectedly, participants identified as using the XYZC rule were well-fit by the Rule-XYZC model in both conditions for both the categorization test (see Figures 9 and 10, panel F) and the feature completion test (see Figures 11 and 12, panel F). By contrast, the fit of the GCM to those same participants was relatively poor; see panel B in Figures 9 through 12. We do not consider the XYZC groups further.

Of greatest remaining interest were the participants who did not use the XYZC strategy (shown in the left-hand panels in Figures 9 through 12); in particular, whether the models can capture the presence of correlational sensitivity in these participants in the probabilistic condition, and its absence in the deterministic condition, and whether that sensitivity is related to the diffusion of attention across multiple dimensions. Because these participants clearly appeared to be using the XOR rule, we hereafter use the term “XOR group” to refer to those people in both conditions. (These were the 18 and 15 participants, respectively, that were assigned to Centroid 2 in the deterministic and probabilistic condition; see Table 7).

As shown in Figure 9 (left column of panels), most of the participants in the deterministic XOR group exhibited a response pattern consistent with the predictions of the GCM and the Rule-XOR model (but not the Rule-XYZC model). The Rule-XOR model provided the best fit to these data based on BIC (see Table 8), although the GCM can quite effectively mimic the Rule-XOR model by shifting attention to the X and Y dimensions (see Table 9). A similar pattern is shown in Figure 10 for the probabilistic condition, although here the XOR pattern was attenuated in the data due to probability matching. Thus, based on the categorization transfer data alone, the GCM and the
Rule-XOR model are difficult to tease apart. In both conditions, the transfer data of XOR participants were fit almost equally well by the Rule-XOR model and by the GCM.

Turning to the feature-prediction data, recall that participants were considered to have responded in a manner commensurate with training if they responded with feature value 0 for items 1 to 4 and with feature value 1 for items 5 to 8 (see Table 1 and the discussion in connection with Experiment 2). The left-hand column of panels in Figures 11 shows that participants in the deterministic XOR group failed to predict the training-congruent missing feature. Nonetheless, instead of responding randomly, participants in this group applied a consistent strategy: If the value of X was equal to the value of Y, those participants tended to respond with one color; if X and Y were different, they responded with the other color (see Figure 11 and Table 1). Because each X and Y pattern was shown with both colors during training, none of the models accurately predicted performance of XOR participants in the deterministic condition.

For the probabilistic condition, by contrast, there was a clear trend in the XOR group to respond with the missing feature as expected on the basis of training; furthermore, only the GCM predicted this type of performance (see Figure 12, panel A). The superior fit of the GCM suggests that the correlational sensitivity revealed by the feature prediction task in the probabilistic condition was driven by memory for the training items. Turning to the parameter values from the GCM, it is also evident that the XOR group in the probabilistic condition was characterized by a broader attention profile than in the deterministic condition (see Table 9). Taken together, those two aspects of the GCM’s performance provide a theoretical link in support of our central thesis, namely that the diffusion of attention across multiple dimensions underlies correlational sensitivity when present.

Turning to the normalized attention weights from the two best-fitting models (i.e., the Rule-XOR model for the deterministic condition and the GCM for the probabilistic
condition), the deterministic XOR group focused solely on the X and Y dimensions for the categorization transfer tests ($w_X = 0.64, w_Y = 0.36, w_Z = 0, w_C = 0$) and additionally on the category label for the feature prediction test ($w_X = 0.41, w_Y = 0.23, w_Z = 0, w_{cat} = 0.36$). By contrast, in the probabilistic condition, attention was spread across all of the stimulus dimensions in both tasks (categorization transfer test, $w_X = 0.31, w_Y = 0.14, w_Z = 0.25, w_C = 0.3$; feature prediction test, $w_X = 0.34, w_Y = 0.17, w_Z = 0.26, w_{cat} = 0.23$).

Summary of Modeling Results

The modeling results are readily summarized. (1) Examination of fits to individual participants suggested the presence of different subgroups who utilized different compound cues. In consequence, we fit the three candidate models—the GCM, the Rule-XOR, and the Rule-XYZC variant of the SRM—separately to the two principal groups in each condition. (2) Not unexpectedly, the XYZC group in both conditions was best accommodated by the Rule-XYZC model. This result confirms that a subset of participants elected to utilize a rather complex rule, but beyond that it is of little theoretical interest in the current context. (3) The remaining participants who did not use the XYZC compound relied on the XOR rule formed by the valid XY compound in both conditions. (4) Crucially, notwithstanding their common utilization of XY, the groups differed drastically between conditions with respect to their correlational sensitivity: Whereas the deterministic XOR group exhibited no knowledge of the non-diagnostic $zc$ correlation—thus being modeled best by the Rule-XOR model—the probabilistic XOR group demonstrated clear knowledge of $zc$ on the feature-prediction test. (5) This correlational sensitivity could only be captured by the GCM, and not the two rule models, suggesting that storage of more than just a rule is required to permit the emergence of correlational knowledge. (6) The pattern of parameter estimates in the GCM supported
the contention that a broader attention profile is required to model cor{}
sential sensitivity.

**General Discussion**

**Summary of Results**

The present article offers several contributions: (1) Experiments 1 and 2 replicated the decreased utilization of a decreased validity cue (compare the deterministic and the probabilistic variants; Edgell, 1978, 1980; Edgell et al., 1996). That is, probabilistic feedback reduced the validity of the valid cues; accordingly, utilization rates in the probabilistic condition were also decreased. (2) Additionally, the deterministic conditions of Experiment 2 replicated the well-established finding that people are insensitive to non-diagnostic correlations among cues in category learning (Chin-Parker & Ross, 2002; Medin et al., 1982; Wattenmaker, 1991, 1993). (3) The current experiments extended previous work (see e.g., Medin et al., 1982; Wattenmaker, 1993) by showing that participants favored conjunctions of cues that were of the same type (i.e., Y and Z), in preference to conjunctions of cues of different types (i.e., Y and C). This result highlights the fact that relational similarity between cues contributed to participants' decisions. (4) The data showed that with probabilistic feedback, a large sub-group of people (around 42%) were sensitive to non-diagnostic correlational information. To our knowledge, this result represents the first demonstration of correlational sensitivity in intentional category learning when prior knowledge was not available to guide learning. The fact that not all of the participants demonstrate correlational sensitivity may be a consequence of the over-arching XYZC component which absorbs sensitivity to zc. The validity of XYZC cannot be avoided without introducing another overlapping valid component as in Experiment 1. Nevertheless, the aggregate analysis of Experiment 2 clearly demonstrates that participants were sensitive to a non-diagnostic correlation.
(5) The modeling of Experiment 2 went beyond previous research by demonstrating that (a) a diffusion of attention across multiple cues is central to the acquisition of correlational knowledge and that (b) this correlational knowledge is best captured by an exemplar-based model and may not be readily explainable by a rule-based model. Because participants were not given the instructions to the feature completion task until immediately prior to its commencement, the representations used on the feature completion test had to be acquired at. Hence, we suggest that probabilistic feedback leads to exemplar storage which allows access to correlated cues via selective attention.

Connections to Prior Research

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

Another interesting case of correlational sensitivity in intentional category learning is knowledge partitioning. In knowledge partitioning, training typically involves stimuli comprised of one or two continuous dimensions and a binary “context” dimension (often
Correlated Cues in Probabilistic Categorization

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

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

**Theoretical Implications**

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

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


Author Note

Preparation of this article was facilitated by several Discovery Grants from the Australian Research Council and an Australian Professorial Fellowship to the second author. The first author was supported by a Jean Rogerson post-graduate scholarship. The authors would like to thank John Kruschke and two anonymous reviewers for commenting on an initial draft of this manuscript.

Address correspondence to the first author at Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405. Electronic mail may be sent to daniel.r.little@gmail.com. Personal home page at http://www.cogsciwa.com/.
Footnotes

1 Attention-shifting in probabilistic environments is often assumed to attenuate over the course of learning as participants become accustomed to an unavoidable level of error (Kruschke & Johansen, 1999). At first glance, this attenuation might appear to militate against the detection of non-diagnostic correlated features. However, that attenuation has been shown to be far from complete and participants can learn relevance shifts in a probabilistic task even after 300 trials (Craig, Lewandowsky, & Little, 2008). Hence, we do not consider attenuation to be problematic in the current experiments.

2 The reader may have noticed that the category space also included a non-diagnostic correlation, involving the shading of one of the circles (Z) and color (C); see Table 1. This non-diagnostic correlation is a necessary consequence of having two identically-valid diagnostic components; however, this correlation is of lesser interest in this study because it is difficult to isolate its role from the contribution of the other pairwise correlations (YZ and YC).

3 Of course, in the deterministic condition, probability matching to the objective deterministic probabilities is identical to maximizing. The only other option is undershooting which would represent a failure to learn the task.

4 The data from 7 participants in the deterministic condition was lost due to computer error; hence, the similarity analysis used data from 14 participants in that condition.

5 Hereafter, we introduce the mnemonic of referring to valid components in upper case (e.g., XY and XYZC) and non-valid correlated cue of interest in italics and lower case (e.g., zc).

6 Due to computer error, two participants in the deterministic condition and three participants in the probabilistic condition only received one repetition of the transfer items. Hence, the transfer analysis could not be conducted for these participants.
If it is instead assumed that each item contributes to feature storage along with its corresponding category label, the fits of the models remain unchanged.
Table 1

*Training stimuli and transfer items for Experiments 1 and 2.*

<table>
<thead>
<tr>
<th>Stimulus Features&lt;sup&gt;a&lt;/sup&gt;</th>
<th>P(A) Feedback</th>
<th>Experiment 1</th>
<th>Experiment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Deterministic</td>
<td>Probabilistic</td>
</tr>
<tr>
<td>CN1 0 0 0 0 0</td>
<td>1.00</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td>CN2 1 0 0 0 0</td>
<td>1.00</td>
<td>0.75</td>
<td>0.00</td>
</tr>
<tr>
<td>CN3 0 1 0 0 0</td>
<td>0.00</td>
<td>0.25</td>
<td>0.00</td>
</tr>
<tr>
<td>CN4 1 1 0 0 0</td>
<td>0.00</td>
<td>0.25</td>
<td>1.00</td>
</tr>
<tr>
<td>CN5 0 0 1 1 0</td>
<td>0.00</td>
<td>0.25</td>
<td>1.00</td>
</tr>
<tr>
<td>CN6 1 0 1 1 1</td>
<td>0.00</td>
<td>0.25</td>
<td>0.00</td>
</tr>
<tr>
<td>CN7 0 1 1 1 1</td>
<td>1.00</td>
<td>0.75</td>
<td>0.00</td>
</tr>
<tr>
<td>CN8 1 1 1 1 1</td>
<td>1.00</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td>IN1 0 0 0 1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IN2 1 0 0 1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IN3 0 1 0 1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IN4 1 1 0 1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IN5 0 0 1 0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IN6 1 0 1 0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IN7 0 1 1 0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IN8 1 1 1 0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<sup>a</sup>The physical instantiation of the three circle features was randomly allocated to the abstract X, Y, and Z dimensions for each participant.

<sup>b</sup>The prefix of the stimulus code refers to an items status at transfer with CN meaning consistent and IN meaning inconsistent.
Table 2

*Base rate and stimulus component validity for Experiments 1 and 2.*

<table>
<thead>
<tr>
<th>Stimulus Components</th>
<th>Base Rate</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>C</th>
<th>XY</th>
<th>XZ</th>
<th>XC</th>
<th>YZ</th>
<th>YC</th>
<th>ZC</th>
<th>XYZ</th>
<th>XYC</th>
<th>XZC</th>
<th>YZC</th>
<th>XYZC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deterministic</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Probabilistic</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Experiment 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deterministic</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>Probabilistic</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Table 3

*Stored feature-to-category associations for the Rule-XOR model.*

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

*Missing features are indicated by a hyphen.*
Table 4

*Stored feature-to-category associations for the Rule-XYZC model.*

<table>
<thead>
<tr>
<th>Category</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 5

*Individual model fits (-lnL) to the categorization and feature prediction tests of the deterministic feedback condition from Experiment 2.*

<table>
<thead>
<tr>
<th>Participant</th>
<th>Model</th>
<th>Model</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GCM</td>
<td>Rule-XOR</td>
<td>Rule-XYZC</td>
</tr>
<tr>
<td>1</td>
<td>110.21</td>
<td>110.97</td>
<td>198.36</td>
</tr>
<tr>
<td>2</td>
<td>192.91</td>
<td>193.50</td>
<td>194.54</td>
</tr>
<tr>
<td>3</td>
<td>160.42</td>
<td>199.61</td>
<td>25.20</td>
</tr>
<tr>
<td>4</td>
<td>77.82</td>
<td>88.72</td>
<td>198.86</td>
</tr>
<tr>
<td>5</td>
<td>197.86</td>
<td>199.43</td>
<td>167.20</td>
</tr>
<tr>
<td>6</td>
<td>186.46</td>
<td>186.46</td>
<td>154.38</td>
</tr>
<tr>
<td>7</td>
<td>128.76</td>
<td>128.79</td>
<td>199.17</td>
</tr>
<tr>
<td>8</td>
<td>98.49</td>
<td>99.47</td>
<td>199.24</td>
</tr>
<tr>
<td>9</td>
<td>110.97</td>
<td>110.97</td>
<td>196.54</td>
</tr>
<tr>
<td>10</td>
<td>160.67</td>
<td>199.61</td>
<td>16.56</td>
</tr>
<tr>
<td>11</td>
<td>99.48</td>
<td>99.47</td>
<td>199.58</td>
</tr>
<tr>
<td>12</td>
<td>94.78</td>
<td>94.80</td>
<td>199.60</td>
</tr>
<tr>
<td>13</td>
<td>164.34</td>
<td>164.45</td>
<td>199.47</td>
</tr>
<tr>
<td>14</td>
<td>179.38</td>
<td>199.43</td>
<td>164.07</td>
</tr>
<tr>
<td>15</td>
<td>137.37</td>
<td>199.01</td>
<td>129.95</td>
</tr>
</tbody>
</table>
Table 6

*Individual model fits (-lnL) to the categorization and feature prediction tests of the probabilistic feedback condition from Experiment 2.*

<table>
<thead>
<tr>
<th>Participant</th>
<th>Model</th>
<th>GCM</th>
<th>Rule-XOR</th>
<th>Rule-XYZC</th>
<th>Participant</th>
<th>Model</th>
<th>GCM</th>
<th>Rule-XOR</th>
<th>Rule-XYZC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>111.80</td>
<td>142.99</td>
<td>199.59</td>
<td>15</td>
<td>1</td>
<td>139.98</td>
<td>180.18</td>
<td>199.19</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>128.28</td>
<td>120.67</td>
<td>182.43</td>
<td>16</td>
<td>2</td>
<td>109.73</td>
<td>181.34</td>
<td>199.57</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>178.14</td>
<td>199.29</td>
<td>73.29</td>
<td>17</td>
<td>3</td>
<td>87.65</td>
<td>104.74</td>
<td>199.57</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>196.52</td>
<td>199.08</td>
<td>193.50</td>
<td>18</td>
<td>4</td>
<td>141.19</td>
<td>156.85</td>
<td>199.48</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>190.04</td>
<td>199.16</td>
<td>52.01</td>
<td>19</td>
<td>5</td>
<td>192.11</td>
<td>196.10</td>
<td>197.86</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>197.02</td>
<td>197.59</td>
<td>198.93</td>
<td>20</td>
<td>6</td>
<td>189.45</td>
<td>188.89</td>
<td>191.49</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>193.97</td>
<td>193.63</td>
<td>119.41</td>
<td>21</td>
<td>7</td>
<td>190.53</td>
<td>188.22</td>
<td>149.05</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>196.57</td>
<td>196.42</td>
<td>199.19</td>
<td>22</td>
<td>8</td>
<td>198.67</td>
<td>199.63</td>
<td>147.08</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>118.66</td>
<td>199.63</td>
<td>158.99</td>
<td>23</td>
<td>9</td>
<td>191.00</td>
<td>199.63</td>
<td>151.85</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>199.52</td>
<td>197.42</td>
<td>197.42</td>
<td>24</td>
<td>10</td>
<td>199.01</td>
<td>199.60</td>
<td>185.46</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>179.62</td>
<td>195.85</td>
<td>146.03</td>
<td>25</td>
<td>11</td>
<td>178.77</td>
<td>189.98</td>
<td>185.85</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>193.58</td>
<td>199.60</td>
<td>62.42</td>
<td>26</td>
<td>12</td>
<td>196.97</td>
<td>197.44</td>
<td>192.89</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>193.22</td>
<td>196.81</td>
<td>195.64</td>
<td>27</td>
<td>13</td>
<td>176.61</td>
<td>170.39</td>
<td>195.56</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>161.57</td>
<td>191.89</td>
<td>199.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 7

Correlations between \( k \)-means centroids from both the deterministic and probabilistic conditions and the ideal XYZC and \( zc \) response profiles in Experiment 2.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Ideal XYZC</td>
<td>-</td>
<td>0.00</td>
<td>0.96**</td>
<td>-0.12</td>
<td>-0.29</td>
<td>0.76**</td>
</tr>
<tr>
<td>2. Ideal ( zc )</td>
<td>-</td>
<td>-0.01</td>
<td>-0.12</td>
<td>0.90**</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>3. Deterministic Centroid 1 (( N = 11 ))</td>
<td>-</td>
<td>-0.10</td>
<td>-0.30</td>
<td>0.72**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Deterministic Centroid 2 (( N = 18 ))</td>
<td>-</td>
<td>-0.07</td>
<td>-0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Probabilistic Centroid 1 (( N = 13 ))</td>
<td>-</td>
<td>-0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Probabilistic Centroid 2 (( N = 15 ))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** \( p < .01 \)**
Table 8

*Aggregated model fits (BIC’s\(^a\) with RMSD’s listed in parentheses) to the XOR and XYZC groups from the deterministic and probabilistic conditions of Experiment 2.*

<table>
<thead>
<tr>
<th></th>
<th>GCM</th>
<th>Rule-XOR</th>
<th>Rule-XYZC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BIC</td>
<td>RMSD</td>
<td>BIC</td>
</tr>
<tr>
<td>Deterministic XOR</td>
<td>5937</td>
<td>(0.12)</td>
<td><strong>5819.4</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8034.5</td>
</tr>
<tr>
<td>Deterministic XYZC</td>
<td>3977.1</td>
<td>(0.24)</td>
<td>4236.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>3071.5</strong></td>
</tr>
<tr>
<td>Probabilistic XOR</td>
<td><strong>6291.4</strong></td>
<td>(0.12)</td>
<td>6704.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6863.3</td>
</tr>
<tr>
<td>Probabilistic XYZC</td>
<td>5119</td>
<td>(0.15)</td>
<td>5125.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>4470.7</strong></td>
</tr>
</tbody>
</table>

\(^a\) The best-fitting model for each group is identified in bold.
Table 9

Average model parameters (and standard deviations) for the XOR and XYZC groups from the deterministic and probabilistic conditions of Experiment 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>Group</th>
<th>$\gamma$</th>
<th>$r$</th>
<th>$w_X$</th>
<th>$w_Y$</th>
<th>$w_Z$</th>
<th>$w_C$</th>
<th>$w_{Cat}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCM</td>
<td>D-XOR</td>
<td>34.32 (40.34)</td>
<td>1.66 (0.38)</td>
<td>0.65 (0.32)</td>
<td>0.58 (0.37)</td>
<td>0.08 (0.22)</td>
<td>0.36 (0.38)</td>
<td>0.46 (0.44)</td>
</tr>
<tr>
<td></td>
<td>D-XYZC</td>
<td>16.21 (6.75)</td>
<td>0.62 (0.9)</td>
<td>0.52 (0.33)</td>
<td>0.42 (0.35)</td>
<td>0.42 (0.37)</td>
<td>0.26 (0.25)</td>
<td></td>
</tr>
<tr>
<td>P-XOR</td>
<td></td>
<td>16.04 (17.28)</td>
<td>1.4 (0.8)</td>
<td>0.53 (0.38)</td>
<td>0.25 (0.36)</td>
<td>0.31 (0.38)</td>
<td>0.29 (0.28)</td>
<td>0.25 (0.36)</td>
</tr>
<tr>
<td>P-XYZC</td>
<td></td>
<td>4.84 (3.93)</td>
<td>1.05 (0.86)</td>
<td>0.57 (0.36)</td>
<td>0.14 (0.29)</td>
<td>0.19 (0.3)</td>
<td>0.45 (0.4)</td>
<td>0.58 (0.34)</td>
</tr>
<tr>
<td>Rule-XOR</td>
<td>D-XOR</td>
<td>42.91 (82.77)</td>
<td>1.61 (0.45)</td>
<td>0.41 (0.31)</td>
<td>0.24 (0.27)</td>
<td>0.28 (0.33)</td>
<td>0.3 (0.27)</td>
<td>0.4 (0.28)</td>
</tr>
<tr>
<td></td>
<td>D-XYZC</td>
<td>12.04 (8.76)</td>
<td>1.23 (0.38)</td>
<td>0.44 (0.35)</td>
<td>0.44 (0.25)</td>
<td>0.21 (0.29)</td>
<td>0.41 (0.32)</td>
<td>0.45 (0.34)</td>
</tr>
<tr>
<td>P-XOR</td>
<td></td>
<td>13.08 (17.33)</td>
<td>1.14 (0.81)</td>
<td>0.51 (0.35)</td>
<td>0.36 (0.28)</td>
<td>0.19 (0.33)</td>
<td>0.13 (0.24)</td>
<td>0.38 (0.4)</td>
</tr>
<tr>
<td>P-XYZC</td>
<td></td>
<td>3.6 (2.97)</td>
<td>1.57 (0.73)</td>
<td>0.42 (0.39)</td>
<td>0.42 (0.33)</td>
<td>0.09 (0.17)</td>
<td>0.15 (0.29)</td>
<td>0.65 (0.33)</td>
</tr>
<tr>
<td>Rule-XYZC</td>
<td>D-XOR</td>
<td>47.71 (82.25)</td>
<td>0.9 (0.66)</td>
<td>0.47 (0.37)</td>
<td>0.39 (0.35)</td>
<td>0.35 (0.27)</td>
<td>0.3 (0.36)</td>
<td>0.29 (0.27)</td>
</tr>
<tr>
<td></td>
<td>D-XYZC</td>
<td>91.76 (34.73)</td>
<td>0.88 (0.8)</td>
<td>0.24 (0.3)</td>
<td>0.27 (0.29)</td>
<td>0.17 (0.18)</td>
<td>0.27 (0.28)</td>
<td>0.29 (0.37)</td>
</tr>
<tr>
<td>P-XOR</td>
<td></td>
<td>61.92 (56.52)</td>
<td>1.19 (0.78)</td>
<td>0.31 (0.43)</td>
<td>0.29 (0.37)</td>
<td>0.22 (0.33)</td>
<td>0.26 (0.38)</td>
<td>0.35 (0.45)</td>
</tr>
<tr>
<td>P-XYZC</td>
<td></td>
<td>93.05 (125.37)</td>
<td>1.37 (0.74)</td>
<td>0.3 (0.38)</td>
<td>0.28 (0.3)</td>
<td>0.26 (0.28)</td>
<td>0.17 (0.31)</td>
<td>0.27 (0.34)</td>
</tr>
</tbody>
</table>

aThe attention parameters listed in the table are given at their values before normalization.

bGroup names have been abbreviated as follows: D-XOR, Deterministic XOR group; D-XYZC, Deterministic XYZC group; P-XOR, Probabilistic XOR group; P-XYZC, Probabilistic-XYZC.
Figure Captions

Figure 1. Example of the eight training stimuli used in Experiments 1 and 2. The stimuli on the left are shown in green and the stimuli on the right are shown in red.

Figure 2. Mean probability matching scores and 95% confidence intervals for the deterministic and probabilistic conditions of A) Experiment 1 and B) Experiment 2 (see text for details). Dotted lines indicate chance performance (at $PM = .50$) and probability matching (at $PM = .75$).

Figure 3. Aggregate training performance during the final two training blocks (with 95% confidence intervals) and objective probability targets for A) the deterministic condition and B) the probabilistic condition of Experiment 1.

Figure 4. Aggregate transfer utilization rates for each stimulus component (with 95% confidence intervals) for A) the deterministic condition and B) the probabilistic condition of Experiment 1.

Figure 5. Average difference (and 95% confidence intervals) in similarity ratings for all possible stimulus components in Experiment 1. See text for more details. The horizontal line indicates the point of no difference.

Figure 6. Aggregate training data (with 95% confidence intervals) and objective probability targets for A) the deterministic condition and B) the probabilistic condition from Experiment 2.

Figure 7. Aggregate transfer utilization rates for each stimulus component (with 95% confidence intervals) for A) the deterministic condition and B) the probabilistic condition of Experiment 2.
Figure 8. Distribution of average feature completion scores for A) the deterministic condition and B) the probabilistic condition of Experiment 2.

Figure 9. Aggregated model predictions to the deterministic transfer data (means and standard error bars are shown) from Experiment 2. A) GCM fit to the XOR group, B) GCM fit to the XYZC group, C) Rule-XOR fit to the XOR group, D) Rule-XOR fit to the XYZC group, E) Rule-XYZC fit to the XOR group, F) Rule-XYZC fit to the XYZC group.

Figure 10. Aggregated model predictions to the probabilistic transfer data (means and standard error bars are shown) from Experiment 2. A) GCM fit to the XOR group, B) GCM fit to the XYZC group, C) Rule-XOR fit to the XOR group, D) Rule-XOR fit to the XYZC group, E) Rule-XYZC fit to the XOR group, F) Rule-XYZC fit to the XYZC group.

Figure 11. Aggregated model predictions to the deterministic feature prediction data (means and standard error bars are shown) from Experiment 2. A) GCM fit to the XOR group, B) GCM fit to the XYZC group, C) Rule-XOR fit to the XOR group, D) Rule-XOR fit to the XYZC group, E) Rule-XYZC fit to the XOR group, F) Rule-XYZC fit to the XYZC group.

Figure 12. Aggregated model predictions to the probabilistic feature prediction data (means and standard error bars are shown) from Experiment 2. A) GCM fit to the XOR group, B) GCM fit to the XYZC group, C) Rule-XOR fit to the XOR group, D) Rule-XOR fit to the XYZC group, E) Rule-XYZC fit to the XOR group, F) Rule-XYZC fit to the XYZC group.
Correlated Cues in Probabilistic Categorization, Figure 3

Target
Data

Training Items

P(A)

A

B

CN1, CN2, CN3, CN4, CN5, CN6, CN7, CN8

Training Items

P(A)
Correlated Cues in Probabilistic Categorization, Figure 6
Correlated Cues in Probabilistic Categorization, Figure 7
Correlated Cues in Probabilistic Categorization, Figure 9
Correlated Cues in Probabilistic Categorization, Figure 11
Chapter 7

Beyond non-utilization: Irrelevant cues can gate learning in probabilistic categorization

*Paper 2 (Refereed)*

Beyond Non-Utilization: Irrelevant Cues Can Gate Learning in

Probabilistic Categorization

Daniel R. Little and Stephan Lewandowsky

University of Western Australia

Running Head: Irrelevant Cues in Probabilistic Categorization

Address correspondence to: Daniel R. Little

School of Psychology (M304)
University of Western Australia
Crawley, W.A. 6009, AUSTRALIA
littld02@student.uwa.edu.au

Abstract

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 which has shown that people sometimes use irrelevant cues to gate access to partial knowledge encapsulated in independent partitions. We report two 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 partition their knowledge on the basis of an irrelevant cue. We show by computational modeling that knowledge partitioning cannot be accommodated by two exemplar models (GCM and RASHNL) whereas a rule-based model (GRT) can capture partitioned performance. We conclude by pointing to the necessity of a mixture-of-experts approach to capture performance in MCPL and identifying reduction of complexity as a possible explanation for partitioning.
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 mimic probability matching by deterministically responding in the presence of perceptual noise (Ashby & Gott, 1988; Ashby & Lee, 1991). 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., 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 model (Rapid Attention Shift ‘N Learning; Kruschke & Johansen, 1999), 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., GCM; McKinley & Nosofsky, 1995), as well as to rule-based models (e.g., GRT; Ashby & Gott, 1988) which 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 re-evaluated 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” framework. We then present two experiments that confirm the existence of knowledge partitioning in MCPL, which constitutes a counter-intuitive 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 towards 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 sub-components of knowledge in different circumstances (e.g., Yang & Lewandowsky, 2003, 2004). Here, we are concerned primarily with the latter finding, known as knowledge partitioning.

The knowledge partitioning 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. Knowledge partitioning has been shown to arise with experts in domain-relevant tasks (Lewandowsky & Kirsner, 2000), with non-experts in function learning (Kalish, Lewandowsky, & Kruschke, 2004; Lewandowsky et al., 2002), and
with non-experts 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 a 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 (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 (ATRIUM; Erickson & Kruschke, 1998).

Overall, knowledge partitioning 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 or not 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, \( P(A \mid \text{Context}) = P(A) \), and context also did not alter the predictiveness of the remaining relevant cues; \( P(A \mid \text{Context, Relevant Cues}) = P(A \mid \text{Relevant Cues}) \). 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 knowledge partitioning 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 knowledge partitioning 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 knowledge partitioning 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 knowledge partitioning (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 knowledge partitioning.

That said, other evidence is suggestive of the possible occurrence of knowledge partitioning 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 non-predictive cue to partition their knowledge.

Should knowledge partitioning 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
acknowledge the simultaneous co-existence of several partial knowledge components. None of the extant theories are likely to meet those requirements.

We now present two experiments that explored whether knowledge partitioning 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 knowledge partitioning can occur in an MCPL task. The experiment 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 knowledge partitioning, 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 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 non-linearly.

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 knowledge partitioning would be manifest if people generalized along the relevant partial 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, two items outside the trained range of the continuous dimension were introduced at transfer (see Table 1).

Method

Participants

Twenty University of Western Australia undergraduate psychology students 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 which varied in shading and color. Shadings were created using Microsoft Words’ 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 8 blocks of trials, each involving five presentations of the 8 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, a 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, the RMSD (root mean squared deviation) was computed 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 ($PM$) score which summarizes performance relative to the actual training probabilities for each participant:

$$PM_i = (P(A|j) - R_i(A|j)) \times SI_j,$$

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 $PM = -.25$ indicating perfect maximizing) and a positive score
indicates undershooting of the training probabilities (with $PM = .25$ indicating chance performance and $PM > .25$ indicating a reversal of the training probabilities).

Table 2 shows that the average $PM$ score tended towards 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.

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 multi-level regression analysis with shading as the single predictor. A multi-level regression permits analysis of all data without confounding between- and within-subject 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 knowledge partitioning 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 sub-groups 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-cluster squared Euclidean distance while minimizing within-cluster distances. The clusters identified by the analysis corresponded to context-insensitive performance (henceforth,
Irrelevant Cues in Probabilistic Categorization

CI, \(N = 7\); knowledge partitioning (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, independent-samples \(t\)-tests were performed between the two groups using various indicators of training performance. Neither the mean RMSD during training, \(t(11) = -0.73, p > .05\), nor probability matching 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) \(\times\) 10 (Shading) within-subjects ANOVA’s for each group. For the CI group, the analysis only revealed a significant effect of Shading, \(F(9, 54) = 7.08, MSE = 0.18, p < .01, \eta^2_p = .54\). 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 to 7 than for the other items, and also the increased response proportion for item 1 compared to item 2 (accompanied by a lesser upturn for item 10 compared to 9).

By contrast, the KP group modulated the use of shading on the basis of context. The significant Context \(\times\) Shading interaction, \(F(9, 45) = 29.55, MSE = 0.05, p < .01, \eta^2_p = .86\), was accompanied by a significant main effect of Shading, \(F(9, 45) = 2.58, MSE = 0.05, p < .05, \eta^2_p = .34\), and a non-significant effect of Context, \(F(1, 5) = 2.89, p > .05\). The main effect of Shading
was due to items 4 to 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 $F(9, 50) = 23.80, MSE = 0.13, p < .01, \eta^2_p = .81$, but not the CI group, $F(9, 60) = 0.21, p > .05$.

**Discussion**

Experiment 1 sought to determine if 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 sub-group 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 knowledge partitioning was present in another sub-group 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 knowledge partitioning 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 knowledge partitioning was characterized by extrapolation outside the trained probability region. To ensure that the extrapolation observed in the KP group was due to learning and 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 enhance the knowledge partitioning 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 knowledge partitioning 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 knowledge partitioning or context-insensitive performance 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 knowledge partitioning, 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, & Süß, 2005), it may likewise be the case that other variables related to general intelligence predict whether a person partitions their 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 University of Western Australia undergraduate psychology students received partial course credit or remuneration (A$10/hr) 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 to 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 minutes 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 to Experiment 1 (see Table 2).

Transfer Performance

Aggregate Analysis. The success of the context manipulation was again examined by a multi-level regression using the difference scores between contexts as dependent measure and stimulus numerosity as 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 which 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, \( p > .05 \), probability matching 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) \( \times \) 10 (Numerosity) within-subjects ANOVA’s. The significant Context \( \times \) Numerosity, \( F \) (9, 81) = 6.69, \( MSE = .11, p < .01, \eta^2_p = .43 \), interaction in the KP group and the significant main effect of Numerosity, \( F \) (9, 90) = 8.20, \( MSE = .22, p < .01, \eta^2_p = .45 \), 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, \( MSE = .11, p < .05, \eta^2_p = .38 \). The largest of the remaining non-significant 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, \( MSE = .22, p < .01, \eta^2_p = .37 \), but not the CI group, \( F \) (1, 100) = 0.52, \( p > .05 \).

Digit Symbol Rotation

The rotation scores were first aggregated across participants, yielding a mean total number correct (M = 25.16, SD = 9.33) comparable to 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 knowledge partitioning 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 knowledge partitioning should not be found in tasks which 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.

We now seek to pinpoint the underlying representations. If the CI and KP groups are using different strategies and employ 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 Generalized Context Model (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 General Recognition Theory (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, the GCM was found to provide the best fit for the CI group, and 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:

\[ s_{ij} = \exp(-c \cdot d_{ij}), \]  

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:

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

where \( x_{ik} \) is the value of dimension \( k \) for the test item \( i \) and \( x_{jk} \) 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):

\[
P(A | i) = \frac{\left( \sum_{j \in A} M_{jA} \cdot s_{ij} \right)^{\gamma}}{\left( \sum_{j \in A} M_{jA} \cdot s_{ij} \right)^{\gamma} + \left( \sum_{j \in B} M_{jB} \cdot s_{ij} \right)^{\gamma}},
\]  

(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 through 5, and the reverse for items 6 through 9 (with the complement forming the arrays for \( M_{jB} \)).

The response scaling parameter, \( \gamma \), allows responding to vary between probability matching when \( \gamma \approx 1 \) 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 $i$, $\alpha_i$, is determined by a set of underlying gains, $\vartheta$ (the use of gains enables normalization of attention and computation of the derivative of attention with respect to error):

$$
\alpha_i = \exp\left(\vartheta_i\right) / \left(\sum_j \exp(\vartheta_j)\right)^{1/\Gamma}
$$

The summation is over all dimensions and $\Gamma$ is an attention-normalization constant. When $\Gamma$ is set to unity, any increase in attention to one dimension requires an equal decrease in attention to all other dimensions. When $\Gamma$ is greater than one, any increase in attention results in a smaller decrease in attention to the other dimension. The reverse is true when $\Gamma$ 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 $\Lambda$. 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 $i$ ($\varepsilon_i$) is identical
Irrelevant Cues in Probabilistic Categorization

23

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:

\[
a_j^{\alpha} = \exp \left( -c \sum_i \alpha_i \epsilon_i |\psi_j - a_i^{\alpha}| \right),
\]

(6)

where \( c \) is the overall width of the receptive field of that exemplar unit (also known as the specificity), \( a_i^{\alpha} \) is the value of the stimulus on dimension \( i \) and \( \psi_j \) 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 (\( \lambda_A \)), a learning rate for the attention gains (\( \lambda_G \)), a shift parameter for the attention strengths (\( \Lambda \)), the attention normalization constant (\( \Gamma \)), a specificity parameter that determines the extent of generalization between exemplar representations (\( \epsilon \)), and the dimensional saliences (\( \epsilon_d \)). In the present simulations, the salience of context was fixed (\( \epsilon_{\text{ctxt}} = 1.0 \)) but the relative salience of the (quasi-) continuous cues (shading or numerosity), \( \epsilon_d \), 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 modelled by a decay of the learning and shift parameters as follows:

\[
r(t) = 1/(1 + \rho \cdot t),
\]

(7)

where \( r \) is the current readiness of the network to learn, \( t \) the trial number, and \( \rho \) 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, $\phi$, that scales activations into response probabilities. The activity of each category node $k$ is given by:

$$a_{cat}^k = \sum_{ex} w_{cat}^{ex} a_{ex}^j,$$  \hspace{1cm} (8)

where the $w_{cat}^{ex}$ 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:

$$P(A) = \frac{\exp(\phi a_{cat}^A)}{\sum_{cat} \exp(\phi a_{cat}^A)},$$  \hspace{1cm} (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 or not 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
Irrelevant Cues in Probabilistic Categorization

× 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.\(^3\) Second, unlike the learning mechanism in RASHNL, the process by which people place the boundary in GRT remains unspecified (e.g., 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, 1996, Experiment 1) or by considering novel transfer items as well (e.g., Maddox & Ashby, 1993; McKinley & Nosofsky, 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, which 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, \(\mu_i\), and variance, \(\sigma_i^2\). 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:

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

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 using 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 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:

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

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 where 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; based on 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 ($w_{AIC}$). 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).

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 $w_{AIC}$ favors the GRT-integrated over all other models. In order to maximize the diagnosticity of the modeling, goodness-of-fit was assessed 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 non-extrapolation 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.

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 \(w_k = 1\)), 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 $\gamma$ with some moderate variation in the final estimate for the attention devoted to context. Panel D, by contrast, 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 $\gamma$. The large values of $\gamma$ 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.

*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 knowledge partitioning, 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 showed that a single change in a long sequence of probabilistic stimuli substantially altered the predictions of ALCOVE, and RASHNL 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, $\phi$, 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 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 knowledge partitioning, 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 knowledge partitioning group. 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 knowledge partitioning. 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 sub-groups of participants and, second, one might be concerned about the exclusion of people from the analysis on that basis. Our response is twofold. First, we note that the aggregate multi-level 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., in press; 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 knowledge partitioning 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. 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 Knowledge Partitioning

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), one of which is dedicated to use of verbalizable rules. Unlike several previous dissociations depending on whether or not category rules were verbalizable (e.g., Ashby, Ell & Waldron, 2003; Ashby & Waldron, 1999; Maddox & Ashby, 2004; Maddox, Bohil & Ing, 2004), Lewandowsky et al. (2006) showed that
this distinction did not eliminate knowledge partitioning. The present data show that knowledge partitioning 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 knowledge partitioning 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 knowledge partitioning 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 knowledge partitioning 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 knowledge partitioning. 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 knowledge partitioning. Hence, the current results, which show that knowledge partitioning 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, Campbell, Hatsopoulos & Hurwitz, 1989; Myers, Lohmeier, & Well, 1994; Nosofsky, Kruschke, & McKinley, 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” over-reliance 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 of 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 knowledge partitioning.

*Exemplars vs. 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 confusible engendered more rule-based responding (Rouder & Ratcliff, 2004). Our studies, by contrast, demonstrated 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 knowledge partitioning did not take the form of a unitary rule representation or a rule-plus-exception
Irrelevant Cues in Probabilistic Categorization

37

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; RULEX; Nosofsky et al., 1994; 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 (2004) 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 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. Inter-dimensional 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 based on a given dimension, as in Shepard's 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 knowledge partitioning. 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., Tennenbaum & 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.
Towards 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 knowledge partitioning 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 knowledge partitioning. 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 (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 knowledge partitioning are negatively correlated with intelligence, postulate the reduction of complexity as a determining factor.
References


Irrelevant Cues in Probabilistic Categorization


Author Note

Preparation of this article was facilitated by several Discovery Grants from the Australian Research Council and an Australian Professorial Fellowship to the second author. The first author was supported by a Jean Rogerson post-graduate scholarship.

Address correspondence to the first author at School of Psychology, University of Western Australia, Crawley, W.A. 6009, Australia. Electronic mail may be sent to littld02@student.uwa.edu.au. Personal home page at http://www.psy.uwa.edu.au/Users web pages/cogscience/.
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).

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.

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 non-zero, 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.
The log-likelihoods in the present case involve a probability density and can therefore exceed unity. In consequence, the $AIC_c$'s which are negated in Equation (11) can be negative.

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) which 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 non-zero 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).

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 knowledge partitioning based on context.

A related function-learning exemplar model (EXAM; DeLosh et al., 1997) can show extrapolation behavior by basing its responses on a local slope estimated from the stored exemplars. However, this model would require modification in order to produce a discrete
response, and, furthermore, Kalish et al. (2004) have demonstrated that EXAM is unable to account for knowledge partitioning in function learning. Hence, we forego further discussion of this model.
Table 1

Stimulus structure used in all experiments

<table>
<thead>
<tr>
<th>Experiment 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Stimulus</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Shading$^a$</td>
<td>0 %</td>
<td>10 %</td>
<td>20 %</td>
<td>30 %</td>
<td>45 %</td>
<td>55 %</td>
<td>70 %</td>
<td>80 %</td>
<td>90 %</td>
</tr>
<tr>
<td>Context</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>P(A)</td>
<td>T$^c$</td>
<td>.2</td>
<td>.4</td>
<td>.6</td>
<td>.8</td>
<td>.8</td>
<td>.6</td>
<td>.4</td>
<td>.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment 2</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Stimulus</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Numerosity$^b$</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Context</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>P(A)</td>
<td>T</td>
<td>.2</td>
<td>.4</td>
<td>.6</td>
<td>.8</td>
<td>.8</td>
<td>.6</td>
<td>.4</td>
<td>.2</td>
</tr>
</tbody>
</table>

$^a$ 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 which are only shown at transfer are marked T.
### Table 2

Average performance measures for the final two training blocks from all experiments

<table>
<thead>
<tr>
<th>Experiment</th>
<th>RMSD M</th>
<th>RMSD SD</th>
<th>PM Score M</th>
<th>PM Score SD</th>
<th>Consistency M</th>
<th>Consistency SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>0.36</td>
<td>0.08</td>
<td>0.00</td>
<td>0.12</td>
<td>0.49</td>
<td>0.44</td>
<td>19</td>
</tr>
<tr>
<td>CI Group</td>
<td>0.38</td>
<td>0.07</td>
<td>-0.05</td>
<td>0.07</td>
<td>0.42</td>
<td>0.51</td>
<td>7</td>
</tr>
<tr>
<td>KP Group</td>
<td>0.40</td>
<td>0.06</td>
<td>-0.05</td>
<td>0.13</td>
<td>0.58</td>
<td>0.37</td>
<td>6</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>0.31</td>
<td>0.10</td>
<td>0.07</td>
<td>0.12</td>
<td>0.51</td>
<td>0.42</td>
<td>36</td>
</tr>
<tr>
<td>CI Group</td>
<td>0.36</td>
<td>0.09</td>
<td>0.01</td>
<td>0.12</td>
<td>0.56</td>
<td>0.33</td>
<td>11</td>
</tr>
<tr>
<td>KP Group</td>
<td>0.29</td>
<td>0.09</td>
<td>0.04</td>
<td>0.13</td>
<td>0.38</td>
<td>0.48</td>
<td>10</td>
</tr>
</tbody>
</table>

Legend: RMSD, Root mean squared deviation; PM Score, Probability Matching Score
Table 3

Model fits\(^a\) 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC(_c) (w(_AIC))</th>
<th>RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training Data</td>
<td>Smooth Transfer</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>KP</td>
</tr>
<tr>
<td>GCM</td>
<td>-38.72 (0.00)</td>
<td><strong>-93.45 (1.00)</strong></td>
</tr>
<tr>
<td>RASHNL</td>
<td>-164.97 (0.00)</td>
<td>-57.58 (0.00)</td>
</tr>
<tr>
<td>GRT-integrated</td>
<td><strong>-181.19 (1.00)</strong></td>
<td>-55.25 (0.00)</td>
</tr>
<tr>
<td>GRT-partitioned</td>
<td>-39.95 (0.00)</td>
<td>-31.28 (0.00)</td>
</tr>
<tr>
<td>Chance Performance</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend: AIC\(_c\), Akaike's Information Criterion corrected for finite samples; w\(_AIC\), Akaike Weights; RMSD, Root mean squared deviation

\(^a\) Minimum AIC\(_c\) among competing models for each group and data set are bold-faced
Table 4

Model fits\(^a\) for combined transfer data from Experiments 1 and 2

<table>
<thead>
<tr>
<th>Model</th>
<th>Smooth Transfer</th>
<th>Raw Transfer</th>
<th>Root mean squared deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\text{AIC}_c)</td>
<td>RMSD</td>
<td>Smooth Transfer</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>KP</td>
<td>CI</td>
</tr>
<tr>
<td>GCM</td>
<td>-101.18 (1.00)</td>
<td>-78.32 (0.01)</td>
<td>-81.96 (1.00)</td>
</tr>
<tr>
<td>RASHNL</td>
<td>-70.39 (0.00)</td>
<td>-66.45 (0.00)</td>
<td>-56.89 (0.00)</td>
</tr>
<tr>
<td>GRT-integrated</td>
<td>-66.99 (0.00)</td>
<td>-</td>
<td>-51.61 (0.00)</td>
</tr>
<tr>
<td>GRT-partitioned</td>
<td>-</td>
<td>-101.11 (0.99)</td>
<td>-</td>
</tr>
</tbody>
</table>

Legend: \(\text{AIC}_c\), Akaike's Information Criterion corrected for finite samples; \(w\text{AIC}\), Akaike Weights; RMSD, Root mean squared deviation

\(^a\) Minimum \(\text{AIC}_c\) among competing models for each group and data set are bold-faced
Figure Captions

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

**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) CI ($N = 7$) and B) KP ($N = 6$) groups.

**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) CI ($N = 11$) and B) KP ($N = 10$) groups.

**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 GCM ($c = 15.21$, $\gamma = 0.79$, $w_k = 0.00011$); Panel (B) for RASHNL ($c = 0.69$, $\Gamma = 0.96$, $\Lambda = 0.26$, $\lambda_A = 0.02$, $\lambda_G = 0.77$, $\phi = 0.96$, $\epsilon_d = 1.00$, $\rho = 0.00000012$); Panel (C) for the integrated GRT with two linear decision bounds (Boundary 1: $-0.01x - 0.45y + 1.97$, Boundary 2: $0.01x + 0.05y - 0.36$, $\sigma^2_{\text{cxt}} = 0.37$, $\sigma^2_d = 4.85$); 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.

**Figure 5.** GCM fits to the combined smooth transfer data from Experiments 1 and 2 for (A) the CI group ($c = 0.66$, $\gamma = 3.00$, $w_k = 0.25$) and (B) the KP group ($c = 33.60$, $\gamma = 100.00$, $w_k = 0.9996$). 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.

**Figure 6.** RASHNL fits to the combined smooth transfer data from Experiments 1 and 2 for (A) the CI group ($c = 2.80$, $P = 0.71$, $\Lambda = 0.00$, $\lambda_A = 0.00032$, $\lambda_G = 0.00$, $\phi = 100$, $\epsilon_d = 0.94$, $\rho = 0.0000015$) and (B) the 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.

Figure 7. Panel A shows the integrated GRT with two linear decision boundaries fit to the combined smoothed transfer data from Experiments 1 and 2 for the CI group (Boundary 1: $0.00x - 0.24y + 0.93$, Boundary 2: $-0.03x + 0.41y - 2.58$, $\sigma_{\text{txt}}^2 = 0.09$, $\sigma_d^2 = 6.04$). Note that $x$ and $y$ refer to the values of context and shading respectively. Panel B shows the partitioned GRT with two linear decision boundaries fit to the combined smoothed transfer data from Experiments 1 and 2 for the KP group ($\beta_1 = 3.64$, $\beta_2 = 5.85$, $\sigma_d^2 = 2.07$). 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 linear boundary.
Figure 1
Figure 2
Figure 3
Figure 4

Irrelevant Cues in Probabilistic Categorization

64
Irrelevant Cues in Probabilistic Categorization

Figure 5
Figure 6
Figure 7
Chapter 8

Toward a clarification of correlated cues in concept learning

*Paper 3*


Manuscript in preparation. University of Western Australia.
Running head: CORRELATED CUE CLARIFICATION

Toward a clarification of correlated cues in concept learning

Daniel R. Little
University of Western Australia

Daniel R. Little
School of Psychology
University of Western Australia
35 Stirling Hwy
Crawley WA 6156
Australia

e-mail: littld02@student.uwa.edu.au
fax: +61 (8) 6488 1006
Abstract

The correlation of cues within a concept or category helps to define the category and connects category members in coherent manner. To date, there is no clear picture of whether people have access to correlational information during concept learning. A large part of this obfuscation is due to the varied constructs that are employed to define what makes a correlated cue across the relevant literature. Furthermore, explaining the existing results requires accounting for interactions between different tasks, prior knowledge and the multitudinous definitions; current theories of correlational knowledge fail to account for these interactions. The current paper advances selective attention as a way to unify the multifarious existing data.
Toward a clarification of correlated cues in concept learning

Feature correlations play several important roles in the representation of concepts. Firstly, the correlation of cues across concepts in part forms systematic boundaries between categories (Chin-Parker & Ross, 2002; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Secondly, the correlation of cues within a concept adds coherency to categories (Malt & Smith, 1984). Accordingly, a substantial amount of research has aimed to determine whether people are sensitive or have access to knowledge about correlated cues in both natural and abstract categories (see e.g., Ahn, Marsh, Luhmann, & Lee, 2002; Chin-Parker & Ross, 2002; Malt & Smith, 1984; Medin, Altom, Edelson, & Freko, 1982; Thomas, 1998; Wattenmaker, 1991, 1993). Knowing whether people have access to correlated cue knowledge is important because access to this information would allow better explanation of current category representations and better prediction of the elements of future category members (Chin-Parker & Ross, 2002). Hence, the benefits that arise from having correlational cue knowledge would make it optimal in light of rational theories of cognition (see e.g., Anderson, 1991a, 1991b). However, a consistent understanding of correlated cues has yet to emerge. Furthermore, the conclusions that can be drawn from this research are often characterized by their contrasting results.

One reason that a consistent picture of correlated-cue knowledge has failed to emerge is due to the use of different methodologies to study whether people can access correlated cue information. Correlated-cue research has been conducted using tasks designed to study concept learning in two different ways. Firstly, in category learning tasks, participants explicitly seek to learn natural or abstract categories by accounting for experimenter provided feedback. Secondly, in category usage tasks, participants make use of existing category knowledge to achieve some auxiliary goal, such as predicting missing features or rating the typicality of category members. In the former type of task, the
results are mixed; sometimes people do have access to the correlated cues (Anderson & Fincham, 1996; Medin et al., 1982; Thomas, 1998) and sometimes they do not (Chin-Parker & Ross, 2002; Thomas, 1998; Wattenmaker, 1993, 1991). By contrast in the latter type of task, the consensus is that people do have access to correlated cue information (Ahn et al., 2002; Chin-Parker & Ross, 2002; McNorgan, Kotack, Meehan, & McRae, 2007; Wattenmaker, 1993, 1991). Another instance when correlational cue knowledge appears to be available is when prior knowledge provides an explicit pointer to correlated information (Hayes, Taplin, & Munro, 1996). This paper provides a brief review of correlated cue research, points to the use of multiple constructs to define correlated cues as the major reason for the lack of agreement among the different methods, and suggests a new way of understanding these results, namely, via selective attention. Finally, we review how selective attention can explain several related correlated cue phenomena.

Much of the incoherence that arises from the correlated cue research is due to the lack of agreement on what defines a correlated cue. At our last count, correlated cue definitions have subsumed three different types of cues relations (e.g., continuous cues, discrete cues, diagnostic cues) and at least three other types of cue relations that deserve a mention here. The correlated cue definitions that have formed the core of the relevant research are as follows: 1) correlation defined as the statistical property of a category distribution (i.e., parameterized in the covariance matrix of a bivariate normal distribution; Ahn et al., 2002; Anderson & Fincham, 1996; Crawford, Huttenlocher, & Hedges, 2006; Huttenlocher, Hedges, & Vevea, 2000; Malt & Smith, 1984; Thomas, 1998); 2) diagnostic between-category feature correlations in which the correlation between features is predictive of category outcome (Medin et al., 1982; Wattenmaker, 1991, 1993, Experiment 3); 3) non-diagnostic within-category feature correlations in which the correlation does not determine category membership but nevertheless provides information about the category members (Chin-Parker & Ross, 2002; Wattenmaker, 1991, 1993). In
Correlated cue clarification

addition to these three definitions that have been explicitly used in correlated cue research, there are three other types of cue relations that should also be considered as part of this canon of research because they involve the correlation of a non-diagnostic cue with a) sets of exemplars (e.g., subtyping; Bott & Murphy, 2007); b) regions of the stimulus space (e.g., knowledge partitioning; (Kalish, Lewandowsky, & Kruschke, 2004; Lewandowsky, Kalish, & Ngang, 2002; Lewandowsky & Kirsner, 2000; Lewandowsky, Roberts, & Yang, 2006; Little & Lewandowsky, in press)); or c) values of another continuous cue (Minda & Ross, 2004). Firstly, we concentrate on the three explicit correlated cue definitions.

Correlated cue definitions

Statistical correlation. Several researchers have opted to define correlated cues as a statistical property of a category distribution Anderson & Fincham, 1996; Crawford et al., 2006; Huttenlocher et al., 2000; Malt & Smith, 1984; Thomas, 1998). For example, one continuous feature can usually take on the full correlational range from negatively to positively correlated (i.e., -1 to +1) with another continuous feature. Typically, the stimuli are drawn from two bivariate normal distributions which represent the different categories (see e.g., Thomas, 1998), though uniform distributions with quasi-continuous cues have also been used (see e.g., Anderson & Fincham, 1996) as have discretely valued cues (see e.g., Malt & Smith, 1984). The different instantiations of this definition underlie different results. Participants show evidence of having access to correlational knowledge if they are trained using category usage tasks which require the specification of the value of one missing continuous feature when given the value of another correlated feature (Crawford et al., 2006). These results contrast with Malt and Smith (1984) who found that correlational information had little effect on participants’ typicality ratings.

In a category learning task, Thomas (1998) trained participants to categorize items drawn from two overlapping bivariate normal distributions that were either positively or
negatively correlated. When the distributions were positively correlated participants demonstrated that they were able to use the correlation between the features to correctly predict the value of one dimension given the value of a second dimension. When the distributions were negatively correlated participants were not able to correctly complete the feature inference task; part of the explanation of the difference between the two conditions was that participants often divided the categories in the negative correlation condition by using a unidimensional boundary (rather than a two-dimensional boundary which was typical of the positive correlation condition) because either type of boundary could be used with equal accuracy (Thomas, 1998). Dividing the categories using a unidimensional boundary meant that participants did not learn about the values of both dimensions and, hence, did not show any sensitivity to the negative correlation between the dimensions. Instead, only knowledge of a single dimension, on which the boundary was defined, was acquired. Hence, vastly different results can arise depending on a) how the statistical correlation is realized and b) the direction of the correlation.

**Diagnostic correlation.** A great deal of the positive evidence of correlational knowledge in category learning comes from tasks where the target correlation is perfectly or partially predictive of category membership (Medin et al., 1982; Wattenmaker, 1991, 1993). In Medin et al. (1982), participants were trained on a simulated medical diagnosis task, which required participants to differentiate two categories on the basis of an XOR cue combination (also called a *biconditional* rule). The XOR combination is relevant if the product of two discrete cue values (coded as -1 and +1) is perfectly predictive of the category outcome. This type of rule can be effectively verbalized as “If both cues have the same value, say Category A; otherwise, say Category B”. Obviously, such a rule is also applicable in tasks which require participants to learn as much as they can about a single category which contains two perfectly correlated features (see e.g., Medin et al., 1982, Experiments 1, 2, and 3; Wattenmaker, 1991, Experiment 1; Wattenmaker, 1993,
Experiment 1). Not surprisingly, when the correlation is diagnostic it is usually the case that participants show sensitivity to it in a variety of different tests (e.g., categorization transfer tests, cue importance ratings, feature inference tests). As noted in Murphy (2002), there is little else to learn in these categories except for the diagnostic correlation. However, it is not always the case that people are sensitive to diagnostic dimensions.

In a series of experiments examining different encoding modes (i.e., incidental vs. intentional encoding; note that these definitions roughly correspond to category usage and category learning as defined above), Wattenmaker (1991) tested for evidence of sensitivity to feature correlations after both types of encoding. In Experiment 1, participants were instructed to find what makes a single category cohere as a category. There were several correlated features including nine XOR correlations and four non-diagnostic correlations; however, sensitivity was tested only for seven of the XOR correlations. The results show that intentional learners show very little sensitivity to the XOR cues (as measured by a two-alternative forced-choice task where participants had to select which category the presented correlation appeared in more often during training). In Experiment 2, the correlations were defined as three-cue probabilistic co-occurrences (that is, certain combinations appear on most but not all trials). Again, the intentional condition was not sensitive to these co-occurrences. In Experiment 3, there were two categories and participants were tested on their sensitivity to three of six XOR cues; there were also five non-diagnostic cues and one imperfect non-diagnostic cue. Transfer tests which pitted typicality, similarity and correlations (XOR) against each other found the intentional learning participants showed very little sensitivity to correlations; instead, typicality was the most important factor in the transfer responses.

Wattenmaker (1993) continued this line of research using a single category with two perfectly co-occurring cues (i.e., the values of the cues were always the same during training) and found that the intentional learning condition was sensitive to this type of
Correlated cue clarification

Correlated cue. Recall that learning a single category with two perfectly co-occurring cues is equivalent to learning an XOR rule (i.e., all non-members of the category will have different values on the correlated dimensions). Experiment 2 tested sensitivity to diagnostic correlated cues (i.e., the values of the cues perfectly predicted the category outcome). For example, fictional creatures from category A were always galloped and barked while creatures from category B always waddled and screeched. Again, participants were sensitive to the correlated features. Experiment 3 also used XOR cues but found decreased sensitivity to the correlated XOR cues, but it is worthwhile to note that family resemblance played a large role in Experiment 3. Although two dimensions formed a perfect XOR rule, several other stimulus components (including single dimensions) had partial validity making rule-plus-exception (and other) strategies possible (Wattenmaker, 1993). In Experiment 4, when correlated features (XOR) were pitted against family resemblance participants showed sensitivity to correlated features over similarity when the items were blocked but not when they were presented in a random fashion; blocking presumably made the correlated features more salient. The results from this rich collection of experiments indicate that if all that is relevant is the XOR correlation and there are a small number of cues then people are sensitive to this correlation; however, if there are a large number of cues or if other cues also have non-zero validities (i.e., other cues are partially predictive of category outcome) then people show little sensitivity to diagnostic correlated features.

Non-diagnostic correlation. Correlated cues have also been defined as within-category feature co-occurrence (Chin-Parker & Ross, 2002). This is similar to defining correlated cues as a statistical correlation but typically refers to the co-occurrence of discrete valued cues rather than continuous cues. For example, in Chin-Parker and Ross (2002), whether a person had a psychology degree or a geology degree was perfectly predictive of whether they worked on Project A or Project B, respectively, but having a
psychology degree was also perfectly correlated with belonging to the racquetball club and having a geology degree was correlated with belonging to the volleyball club. Hence, each item was predictive on its own but the correlation between the features was not predictive. Again, the type of task has a large role in determining whether or not people have knowledge of the feature co-occurrence. For example, Chin-Parker and Ross (2002) examined correlational sensitivity in a feature inference task and a category learning task and found that participants in the inference condition could readily specify the value of a missing cue if given another cue but participants in the categorization condition could not (see also, Kellogg, 1981). Hence, participants in the category learning task were not sensitive to the within category correlation.

In a similar experiment, Billman and Knutson (1996) allowed participants to view at their own pace several items that were either comprised a richly correlated feature structure (i.e., 4 of the features perfectly co-varied) or a single correlated feature structure (i.e., only one pair of items co-varied). Participants in the rich correlation condition were able to correctly identify a missing correlated feature with a high level of accuracy, but participants in the other condition were only able to do so for the pair that was co-varied in their item set. Hence, the sensitivity demonstrated in the feature inference task can be enhanced by using a more complex correlational structure.

Prior knowledge. A further complication is introduced by considering the effects of using natural categories where prior knowledge is available to guide learning rather than abstract categories where all that is available to guide learning is the experimenter controlled reinforcement. For example, Malt and Smith (1984) used natural categories but only found correlational sensitivity for cues that were explicitly noted a priori as being correlated. Ahn et al. (2002) followed up Malt and Smith (1984) by examining the types of properties that were correlated in people’s representations of natural categories. In this experiment, participants tended to note that features that were causally related (i.e., one
feature depends on the other feature, but not vice versa) were correlated. Hence, if participants can explicitly theorize about the correlation (e.g., the bird has the correlated features "flies" and "has wings" because "having wings" allows the bird to "fly") then they will note the correlated features. Features which are not dependent are not noticed as often (e.g., "flying" and "chirping"). Barrett, Abdi, Murphy, and Gallagher (1993) found that children were sensitive to theory-based correlations but not to theory-neutral correlations. In fact, theory-based correlations seem to override learning about other category properties.

Murphy and Wisniewski (1989) trained subjects on family resemblance categories in which a pair of features always co-occurred. Testing on consistent and inconsistent correlation items revealed no effect of the correlations on responding in three experiments. However, when the tested correlations were expected from prior knowledge (e.g., a weapon that "blinds you" also "emits a bright light"), there was evidence that the correlations affected responding, even if the tested features did not occur together during learning any more than other features (Murphy & Wisniewski, 1989; see also Murphy, 2002). Hayes et al. (1996) revisited the experiment used in Wattenmaker (1991) and found that if the categories are given meaningful labels that are consistent with prior knowledge, then participants acquire knowledge of the correlated cues even in the intentional category learning condition. The theoretical implication of these results is that prior knowledge guides learning by providing a pointer to what is or might be relevant. Of course, other explanations are necessary to explain tasks where prior knowledge is not available.

Theoretical explanations

Several explanations exist to try to unify these disparate results. One explanation emphasizes the importance of exemplar storage (Medin et al., 1982; Wattenmaker, 1993, 1991). For example, some tasks promote exemplar storage (e.g., incidental learning tasks).
while some do not (e.g., intentional learning tasks); exemplar storage mediates
correlational sensitivity because stored category instances can be recalled and the
correlational information can be extracted. A prototype or rule would presumably discard
the correlational information. Another explanation posits that the goal of the task
determines whether participants can access correlational information. For example, if the
goal is to correctly learn the criterial attributes of a category, knowledge of non-diagnostic
category elements might be limited (McNorgan et al., 2007). A further explanation is that
people can learn correlations that form part of a coherent structure or in other words, are
consistent with prior knowledge and expectations (Barrett et al., 1993; Hayes et al., 1996;
McNorgan et al., 2007; Murphy & Wisniewski, 1989). While successful in some instances,
none of these explanations are fully capable of explaining all of the above results. We
propose that selective attention provides a mechanism to explain all of the combinations
of prior knowledge present and absent, category learning (i.e., intentional category
learning) and category usage (i.e., feature inference and incidental category learning)
tasks, and the multiple definitions of correlated cues.

Selective Attention. Recall that in Thomas (1998) participants trained on categories
represented by bivariate normal distributions were sensitive to correlated features if the
correlation was positive but not if the correlation was negative. Thomas (1998) proposed
that participants applied a single-dimension boundary instead of a two-dimensional
boundary in the negative correlation condition because they could selectively attend to a
single dimension without a subsequent decrease in accuracy. Hence, selective attention
away from one of the dimensions meant that knowledge of the correlation between the
features was not acquired (Thomas, 1998).

The selective attention hypothesis was explicitly tested by Little and Lewandowsky
(in press) who introduce probabilistic feedback to hypothetically increases attention
shifting by increasing the baseline level of error in the task. ¹ The authors found that
participants in the probabilistic feedback condition (i.e., where category reinforcement was less than perfect and stimuli belonged to multiple categories on a probabilistic basis) showed increased sensitivity to non-diagnostic co-occurring cues compared to a deterministic feedback condition (i.e., a condition where the stimuli always belonged to one category or the other). Furthermore, participants in the probabilistic condition showed a broader attention profile (i.e., stronger attention to multiple dimensions) compared with the deterministic condition. By contrast, in the deterministic condition, attention was narrowly focused on the diagnostic dimensions; hence, prior failures to demonstrate correlational sensitivity in intentional learning tasks (e.g., Chin-Parker & Ross, 2002; Thomas, 1998; Wattenmaker, 1993, 1991 with probabilistic feedback can be explained by a lack of attention to correlated features.

Selective attention can also explain the contrasting positive sensitivity to correlated features shown in category usages tasks. Typically, participants are trained to predict missing features (i.e., a feature inference task; Chin-Parker & Ross, 2002; Crawford et al., 2006; Yamauchi, 1998, 2000) rather than predict a category label or trained to process the stimulus holistically by making likeness ratings (Wattenmaker, 1991, 1993). Attention in these tasks is presumably more diffuse than in category learning tasks; evidence for this conjecture can be found in the related task used by Minda and Ross (2004). In this task, participants were trained to view fictional animals comprised of multiple binary dimensions and a single trinary dimension and either a) make a continuous prediction of how much food that animal would eat or b) categorize the animal first then make a continuous prediction. In the categorization condition, attention (as indexed by feature weightings in a computational model) was focused solely on the diagnostic binary dimension, but in the prediction-only condition attention was spread across all of the stimulus dimensions.

Several alternative proposals have been advanced to explain the role of prior
knowledge in the acquisition of correlational knowledge. One explanation is that prior knowledge is assumed to provide a cache of all previously seen exemplars that can be used in conjunction with the newly acquired exemplars in the current task (Heit, 1993, 1994, 2000). Alternatively, prior knowledge is assumed to propagate exemplars that are consistent with expectations (Heit, 1993, 1998, 2000). The former explanation can explain correlational sensitivity if the repository of stored exemplars contains the appropriate correlational information and this store has a larger influence than newly acquired exemplars (Heit, 2000). However, the attention shifting hypothesis suggests that correlational information is available in category learning when attention is directed to the correlated (but non-diagnostic) features. Hence, another possibility is that prior knowledge guides selective attention to correlated features through direct action on the attention weights. Evidence for this can be derived from studies on subtyping (Bott & Murphy, 2007; Hayes, Foster, & Gadd, 2003), where a within-category correlation marks exceptions items.

The typical finding is that people use the subtyping feature to prevent exceptions from interfering with their prior knowledge congruent category representation. Interestingly, if prior knowledge was not available to the participant before the task began then subtyping effects did not occur; however, when prior knowledge was available, the subtyping effect enhanced reliance on prior knowledge (Bott & Murphy, 2007). A compelling explanation of this effect is that when prior knowledge is available it guides attention to the within-category correlation that marks the exception item; when prior knowledge is not available, no attention is allocated to the correlated feature, and items are not marked as exceptions. Whether this occurs through direct action or selective attention or via some other mechanism (e.g., representational attention; Erickson & Kruschke, 1998, 2002) remains to be tested; however, consideration of related research can inform this query.
Related Correlational Research

Knowledge partitioning. Similar to subtyping, knowledge partitioning involves the use of a within-category correlation to mark particular regions of the stimulus space. In knowledge partitioning experiments, participants are typically trained to categorize stimuli comprised of one or two continuous dimensions and a binary "context" dimension (often instantiated as the color of the stimulus; see e.g., Kalish et al., 2004; Lewandowsky et al., 2002; Lewandowsky & Kirsner, 2000; Lewandowsky et al., 2006; Little & Lewandowsky, in press; Yang & Lewandowsky, 2003, 2004. Context, like the subtyping feature, by itself does not predict category membership. However, context reliably identifies which of a number of partial boundaries on the continuous dimensions are applicable to solve the task. Hence, context provides a within-category correlation with a region of the stimulus space, and can be used to break the problem into simpler components. For instance, a typical knowledge partitioning solution might be described as 'if the stimulus is green, large items belong to category A and small items belong to category B; if the stimulus is red, small items belong to category A and large items belong to category B'. Generally, about one third of participants use context to determine which rules to apply. Extensive modeling by Yang and Lewandowsky (2004) and Little and Lewandowsky (in press) determined that selective attention alone was not sufficient to capture the application of different rules in each context. A gating mechanism was required to adjudicate between different applicable rules on the basis of context (Yang & Lewandowsky, 2004). Similar proposals have been made to combine prior knowledge and current learning in the subtyping paradigm (Hayes et al., 2003; Bott & Murphy, 2007), and subtyping can be considered a specific type of knowledge partitioning where the applicable ‘rules’ are the associative links consistent with prior knowledge and the associative links consistent with current learning. Hence, knowledge partitioning and subtyping can be considered positive demonstrations of correlational knowledge that go beyond selective attention.
Conclusion

Firstly, it is obvious that the multitude of definitions is a mixed blessing to category learning research; on one hand, correlated cues have been examined with a variety of different tests and methods, but on the other hand, the results are difficult to interpret. Hence, we recommend that correlated features should be identified as diagnostic or non-diagnostic features regardless of whether they are discrete or continuous. Using diagnostic correlated features to study correlational knowledge is problematic because knowledge of the correlation is confounded with knowledge of diagnosticity. Prior knowledge is an important factor in determining whether correlational knowledge is available and future research should be directed toward determining whether prior knowledge guides selective attention. Though selective attention can explain correlational knowledge, modifications of selective attention are required to explain more complex correlated cue phenomena (e.g., representational attention and knowledge partitioning).
References


Hayes, B., Foster, K., & Gadd, N. (2003). Prior knowledge and subtyping effects in


Author Note

Preparation of this article was facilitated by a Jean Rogerson post-graduate scholarship to the author.

Address correspondence to Daniel Little at School of Psychology, University of Western Australia, Crawley, W.A. 6009, Australia. Electronic mail may be sent to daniel.r.little@gmail.com. Personal home page at http://www.cogsciwa.com.
Footnotes

1 Attention is hypothesized to briefly but fleetingly shift from the learned diagnostic dimensions to other aspects of the current stimulus when participants make a classification error (Kruschke & Johansen, 1999; Kruschke, 2001). Rapid attention shifting was explicitly proposed for probabilistic tasks to accommodate the multiple interactions between diagnostic and non-diagnostic cues that occur when error cannot be avoided (Kruschke & Johansen, 1999).
Chapter 9

Knowledge and expertise

Knowledge and Expertise

Stephan Lewandowsky and Daniel Little

University of Western Australia

and

Michael L. Kalish

University of Louisiana

Word count: 10,539

Running head: Knowledge and Expertise

Address correspondence to: Stephan Lewandowsky
School of Psychology
University of Western Australia
Crawley, W.A. 6009, AUSTRALIA
lewan@psy.uwa.edu.au

URL: http://www.psy.uwa.edu.au/user/lewan/
1. Knowledge in Applied Settings

What is knowledge? In fact, it may be simpler to ask, what is not knowledge? There is perhaps no single more all-encompassing concept in cognitive psychology than knowledge: Knowledge contributes to simple perceptual tasks such as object recognition, when people identify an ambiguous stimulus on the basis of prior knowledge (e.g., Bar & Ullman, 1996). Knowledge contributes to memory performance in a myriad of ways, for example when people reconstruct events according to a schema or script (e.g., Tuckey & Brewer, 2003; Roediger, Meade, & Bergman, 2001). Finally, knowledge is the fundamental ingredient of human cognition at its best, namely expert performance. Accordingly, the literature on knowledge is vast, and its sheer size prevents a summary by a few simple assertions. To keep our task manageable we therefore imposed some strong constraints on this chapter.

Charness and Schultetus (1999) defined knowledge as “acquired information that can be activated in a timely fashion in order to generate an appropriate response” (p. 61). We accept this as our working definition but additionally restrict consideration to manifestations of knowledge that have been variously called declarative or explicit (Reber & Squire, 1994; Shanks & Johnstone, 1998). These forms of knowledge are characterized by being accessible to awareness and verbal report, for example in response to a query such as “what is the capital of France?” We do not give much consideration to issues of training and knowledge acquisition, which are the domain of the chapter on instruction in this handbook. Finally, we use the applied focus of this handbook to guide which topics to foreground and which to downplay. Accordingly, we omit discussion of
computational models of knowledge and its acquisition and transfer (e.g., ACT; Anderson, 1990). Extensive treatments of those models can be found elsewhere (e.g., Singley & Anderson, 1989). Instead, we foreground research on expertise and expert performance; we focus on the fractionation and encapsulation of knowledge; and we examine the success or failure of the transfer of that knowledge.

We proceed as follows: In the first major section we discuss the nature of expert behavior. In particular, we suggest that expertise is the result of specific learned adaptations to cognitive processing constraints. In consequence, expertise turns out to be very specific and “brittle”; that is, experts may encounter difficulties when tasks are altered or when transfer to new problems is expected. We conclude the section on expertise by examining three short-comings and sources of error that experts frequently encounter. In the second major section, we consider more conventional, non-expert manifestations of knowledge. We begin by considering the widespread view that knowledge is integrated and coherent, exemplified by knowledge space theories as well as the mental model approach. We then consider the alternative position; namely, that knowledge is often fragmented or partitioned, and that multiple alternative pieces of knowledge are often held simultaneously. In the final major section, we consider the mechanisms underlying the transfer of learned knowledge to novel situations. We suggest that transfer succeeds only if people perceive the similarity between their existing knowledge and a novel problem, and we then review the factors that affect the perception of similarity. We conclude the section by examining additional factors that may lead to the failure of transfer. Throughout the chapter, we place particular emphasis on the
problems and shortcomings associated with those processes, because they are of major relevance to the practitioner.

2. Expertise and Its Limitations

We begin by considering the performance of the most skilled of individuals, namely experts. Analysis of expertise can illustrate the essential characteristics of human knowledge; indeed, some have gone as far as to argue that expertise is an indicator of consciousness (Rossano, 2003). Our intention in this section is to provide a fairly atheoretical summary of the performance characteristics and shortcomings of experts. The subsequent sections provide a more theoretical discussion of the properties of knowledge in general, and in so doing also provide another, more theoretical perspective on expertise.

Although the many definitions of an expert include anecdotal descriptions such as “anyone who is holding a briefcase and is more than 50 miles from home” (Salthouse, 1991, p. 286) or “someone who continually learns more and more about less and less” (Salthouse, 1991, p. 286), there is common agreement that an expert is characterized by reproducible superior performance in a particular domain. Any coherent set of tasks and problems that is amenable to objective performance measurement (Ericsson, 1996) can constitute a domain of expertise. Accordingly, researchers have examined domains as diverse as the linking of car crime series by expert investigators (Santtila, Korpela, & Hakkanen, 2004), the ability to predict the spread of bush fires by expert fire fighters (Lewandowsky & Kirsner, 2000), medical diagnosis (e.g., Patel, Kaufman, & Magder, 1996), seemingly mundane but highly sophisticated activities such as driving a car (consult the chapter on driving in this handbook for more details), or the performance of
chess masters (e.g., Charness, Krampe, & Mayr, 1996). In all cases, expert performance has been consistently and reliably found to be outstanding and superior to that of novices.¹

In chess, for example, expertise is associated with an extraordinary ability to remember the location of pieces on a board after a few seconds of viewing time, and the ability to play several games at the same time (e.g., de Groot, 1965). In medical diagnosis, experienced radiologists reliably outperform residents when inspecting X-rays (Norman, Brooks, Coblentz, & Babcock, 1992). A contemporary mnemonicist, Rajan, has memorized the digits of $\pi$ to over 30,000 places and can reproduce sequences of up to 75 digits with great ease (e.g., Thompson, Cowan, & Frieman, 1993). Even relatively mundane tasks such as waiting tables (Ericsson & Polson, 1988) and transcription typing (Salthouse, 1991) can involve rather astonishing levels of knowledge and cognitive performance. The chapter on Cognitive Abilities in this handbook provides more information about some of those feats. Notwithstanding the generally high level of performance, expertise is characterized by several attributes that in addition to supporting exceptional performance also engender intriguing limitations and create the potential for serious error.

2.1 Characteristics of Expertise

Circumventing known processing limitations. Expert performance often seems to defy known human processing limitations. For example, it is known that people cannot tap a finger repetitively more than about 6 times a second, even if they do not have to respond to specific stimuli (Freund, 1983). In conjunction with the known lower limit on response latency to successive stimuli (around 550 ms; Salthouse, 1984), these
Knowledge and Expertise

constraints seem to dictate a maximum transcription typing speed of somewhere between 20 and 75 words per minute. Yet, expert typists can enter text at a rate exceeding 75 words per minute. Salthouse (1984) showed that typists achieve this high level of performance by developing specific strategies to circumvent these processing constraints. Specifically, the maximum typing speed a typist can achieve is correlated with the number of characters that must be simultaneously visible for the typist to maintain their maximum speed. This correlation indicates that growing expertise is associated with increased parallelism of processing, such as that used to pre-plan keystrokes involving opposite hands. One index of this planning is the strong negative correlation between expertise and the delay between keystrokes involving alternate hands, as when “w” is followed by “o”. That is, coordination between the two hands increases with the expertise of a typist. The further fact that the correlation between expertise and inter-key intervals is substantially smaller for repetitions of the same letter—which necessarily involves repeated tapping of the same finger—indicates that expertise often involves the acquisition of skills to circumvent “hard” constraints, rather than a relaxation of those biological or cognitive constraints.

Similarly, outstanding memorial abilities appear to be based on acquired strategies and techniques. To illustrate, consider individuals who gradually raised their digit span by deliberate acquisition of mnemonic techniques. In some particularly dramatic instances, a person’s span increased from the standard 7±2 to 80 or even higher (e.g., Ericsson, Chase, & Faloon, 1980; Staszewski, 1993). These remarkable abilities relied on the acquisition of increasingly larger, richly integrated hierarchical retrieval structures (e.g., Staszewski, 1993), an observation supported by computer simulation (Richman,
Staszewski, & Simon, 1995). Ericsson, Delaney, Weaver, and Mahadevan (2004) recently confirmed that a similar account can capture the immediate memory abilities of the mnemonist Rajan mentioned earlier, notwithstanding earlier opinions to the contrary (Thompson et al., 1993).

The view that expertise represents a learned adaptation to task constraints—as opposed to being the result of innate “talent”—has found a theoretical focus in the work by Anders Ericsson and colleagues (e.g., Ericsson, 2003, 2005). The principal tenet of Ericsson’s view is that expertise arises not from mere prolonged exposure to a task but from extensive “deliberate practice”. Deliberate practice differs from mere exposure and repetition in several important ways: First, deliberate practice involves a well-defined specific task that the learner seeks to master. Second, task performance is followed by immediate feedback. Third, there is opportunity for repetition. Ericsson, Krampe, and Tesch-Römer (1993) provided a very extensive characterization of deliberate practice, and Ericsson (2005) surveys specific instances in which the role of deliberate practice has been established in a variety of expert domains.

There is now considerable consensus that, irrespective of the domain of expertise, 10 years of deliberate practice are required to attain outstanding levels of performance (e.g., Ericsson, 1996). Moreover, experts exhibit some notable commonalities across domains. Table 1 lists some of those commonalities that were identified by Holyoak (1991). The bold-faced entries in the table correspond to issues that we take up in this chapter because we consider them to be particularly critical; the reader is referred to Holyoak (1991) for a discussion of the remainder.
The fact that expertise is the result of very specific adaptations to task demands and processing constraints entails two related consequences: First, experience is typically very specific and limited to the trained domain. Second, expertise is often quite brittle, and seemingly small deviations from a routine task can be associated with surprisingly large performance decrements.

**Specificity of expertise.** It should come as no surprise that expert archaeologists are not necessarily also outstanding oceanographers, and that expert psychologists are unlikely also to be world-class ornithologists. However, the extent of that specificity may exceed the expectations and intuitions of most practitioners. For example, individuals who acquire a phenomenally large digit span after extended training (e.g., Ericsson, Chase, & Faloon, 1980), somewhat soberingly retain the standard limited capacity for other information, namely approximately seven symbols (e.g., Chase & Ericsson, 1981). That is, the same person may struggle to recall “C F G K L P Z” in the correct order while being able to reproduce the sequence “2 9 0 3 4 1 8 9 2 3 0 5 7 1 4 5 2 2 8 1 0” (or indeed an even longer series of digits) flawlessly. Similarly, expert pianists’ acquired ability to tap fingers particularly rapidly (Ericsson et al., 1993) does not generalize to an ability to tap feet at a particularly rapid rate (Keele & Ivry, 1987). Perhaps the most astounding demonstration of specificity is the finding that one year after learning to read text in an unfamiliar transformation (e.g., letters flipped upside down and mirror
reversed), people can re-read pages from a year ago reliably more quickly than new text that is presented in the same transformed script (Kolers, 1976).

**Brittleness of expertise.** A corollary of the specificity of expertise is its “brittleness”; that is, the deterioration in performance that is observed when a domain-relevant task is altered slightly and thus becomes atypical. A classic example of this brittleness involves memory for chess configurations. Chase and Simon (1973) found that an expert chess player could recall the identity and location of pieces on a chess board after fairly brief (5 seconds) exposure with remarkable accuracy. However, this striking ability was limited to plausible configurations that might arise during an actual game. When pieces were randomly arranged, and hence no longer formed a meaningful pattern, the performance of the chess expert deteriorated dramatically. The deterioration of expert memory when domain-relevant stimuli are rendered meaningless by randomization or some other disruption is a fundamental attribute of expertise that has been observed in many domains: A review by Ericsson and Lehmann (1996) cites areas as diverse as the games of bridge, GO, Othello, snooker, basketball, field hockey, volleyball, football, and professional disciplines such as medicine, computer programming, and dance.

Another intriguing aspect of these results, in particular those involving chess, arises out of detailed comparisons between experts and novices. For meaningful game positions, the reproduction skills of chess masters are indubitably far superior to that of novices. For random positions, it used to be a matter of consensus that the expert advantage was completely eliminated. The belief that experts and novices did not differ in their memorial abilities for random board configurations was sufficiently entrenched to be echoed in recent textbooks (e.g., Medin, Ross, & Markman, 2001). However, when
the evidence from numerous studies is considered jointly in a meta-analysis, increasing expertise is found to be associated with a small but clear memory advantage even for random board positions (Gobet & Simon, 1996). This small advantage is most likely due to the experts’ ability to discover even small regularities in otherwise random positions by matching board positions against a repertoire of an estimated 50,000 or so chess patterns stored in long-term memory (Simon & Gilmartin, 1973).

Accordingly, when the degree of randomness (defined by the extent to which basic game constraints are violated) is manipulated, players with greater expertise are better able to exploit any remaining regularities than players with lesser expertise (Gobet & Waters, 2003). The specificity of expertise thus extends to highly subtle regularities indeed.

**Expert transfer.** The characteristics of expertise reviewed in the foregoing should readily generate expectations about how expertise transfers from one task to another: It would seem safe to assume a fair degree of within-domain transfer, albeit perhaps bounded by the observed brittleness of expertise, combined with the likely absence of transfer to tasks outside the expert’s domain.

Indeed, there is considerable support for within-domain transfer. For example, Novick and colleagues (1988; Novick & Holyoak, 1991) showed that mathematical expertise predicts the degree to which solution strategies are transferred from one algebra word problem to another that appears different at the surface but shares the same deep structure. In fact, transfer is observed even when the two problems are presented under two separate experimental cover stories. In one study, the amount of transfer among
experts was found to be up to nine times greater than among novices (Novick, 1988, Experiment 1).

Similarly, in the domain of accounting, Marchant, Robinson, Anderson, and Schadewald (1991) showed that experts in general exhibit significantly more transfer than novices between problems involving the application of taxation laws. An accompanying finding was that when the problems were “anomalous”, that is, constituted exceptions to a general taxation principle, the experts’ subsequent transfer was often reduced to the level shown by novices. Marchant et al. argued that processing of the first exceptional case “…increased the salience of a highly proceduralized strategy that overrides transfer from the analogy in the more experienced group” (p. 283). Thus, while expertise generally facilitates within-domain transfer, it may not do so in cases involving exceptional problems, because experts cannot help but activate their general knowledge even when exceptions to that knowledge must be processed. In consequence, the strongly activated general knowledge may prevent the renewed recognition of an exception to the general knowledge. We revisit the theme of the inevitable activation of expert knowledge below.

Turning to the issue of transfer outside a problem domain, it is unsurprising that such transfer is often absent. What is perhaps more surprising is how little deviation from a routine task it takes in order to eliminate transfer. Sims and Mayer (2002) examined the spatial skills of expert Tetris players. Tetris is a computer game that requires the player mentally to rotate shapes presented on the screen in a limited amount of time. People who were experienced Tetris players (either pre-experimental experts or trained in the experiment) did not differ from novices on a whole battery of spatial tests, with the
highly selective exception of mental rotation tests involving either shapes used in Tetris or very similar ones. That is, even though Tetris relies almost entirely on mental rotation skills, and even though people improved those skills during training, this improvement was narrowly limited to a certain type of stimuli and did not transfer to other shapes.

The characteristics of expertise just reviewed can engender specific performance errors and shortcomings that are worthy of the practitioner’s attention. We next review those errors and shortcomings before examining the knowledge structures that underlie skilled performance in general and expertise in particular.

2.2 Expert errors and shortcomings

There is growing recognition that the analysis of performance errors and limitations contributes in fundamental ways to our understanding of the nature of expert knowledge (e.g., Johnson, Grazioli, Jamal, & Zualkernan, 1992). Holyoak (1991) provided a list of expert limitations that are reproduced in Table 2 together with others identified by ourselves. Three of those limitations and shortcomings—inflexibility, expediency, and mediocrity—are particularly relevant here.

Inflexibility. Inflexibility is revealed when experts are confronted with novel task demands that are inconsistent with their existing knowledge base. In those situations, the need for adaptation may prove to be more challenging to experts than to novices (Frensch & Sternberg, 1989; Sternberg & Frensch, 1992). Using the game of bridge as their domain of expertise, Sternberg and Frensch (1992) examined the effects of various
arbitrary rule changes on the performance of expert and novice bridge players. In general, experts were found to suffer more than novices from any rule change, although the extent of their impairment differed with the type of change. When the rule change involved surface modifications, such as introducing new nonsense names for suits and honour cards, experts suffered less of a performance decrement than when the deep structure of the game was changed, for example by altering the rule determining the opening of each play. The fact that expert disruption was maximal after a change to the deep structure suggests that experts, unlike novices, routinely processed the task at that deep level; a finding that is consonant with much prior research (e.g., Chi, Feltovich, & Glaser, 1981; Dunbar, 1995). Highly skilled performance may thus entail the general cost of reduced flexibility in the face of novel task demands.

In a related vein, Wiley (1998) showed that experts cannot suppress the retrieval of domain-relevant knowledge, even when participants are warned that their knowledge may be inappropriate or misleading in the current task setting. Wiley used a remote association task, in which people have to generate a word that can form a familiar phrase with each one of three presented items. For example, given the stimuli plate, broken, and rest, the word home can be used to form the meaningful phrases home plate, broken home, and rest home. Readers with expertise in baseball may have found this example particularly easy because the target phrase home plate represents a crucial concept in baseball. But what if the stimuli had instead been plate, broken, and shot? The intended word here is glass, although the first two words are compatible with the baseball-consistent completion home. Wiley found that baseball experts, unlike novices, had great difficulty with items that implied—but did not permit—a domain-consistent completion.
The experts’ difficulty persisted even when they were warned beforehand that their domain knowledge would be misleading, suggesting that activation of expert knowledge is automatic and cannot be suppressed.

**Expediency.** Expediency, by contrast, concerns the acquisition phase of expertise, and refers to the fact that experts emphasize efficiency when acquiring a skill. They may, for example, trade knowledge for extended search where many cues could be considered (Charness, 1991; Johnson, 1988). Thus, accumulation of a large knowledge-base allows experts to select the key features of the problem, thereby reducing the number of variables chosen for consideration. An illustrative case of expert expediency was reported by Lewandowsky and Kirsner (2000), who asked experienced wild fire commanders to predict the spread of simulated wild fires. The spread of wild fires is primarily determined by two physical variables: fires tend to spread with the wind and uphill. It follows that with light downhill winds, the outcome depends on the relative strengths of the competing predictors. If wind is sufficiently strong, the fire spreads downhill with the wind, whereas if the wind is too light, the fire spreads uphill against the wind. Lewandowsky and Kirsner found that, with an intriguing exception that we discuss in a later section, experts completely ignored slope and based their predictions entirely on wind. While this gave rise to correct predictions in most circumstances, any fire in which light winds were over-ridden by a strong slope was systematically mis-predicted.

**Mediocrity.** Finally, imperfect expert performance has also been associated with situations in which probabilistic cues must be used to predict uncertain outcomes. For example, when predicting the likely success of applicants to medical school from their prior record (e.g., grades, letters of recommendation), expert accuracy is often inferior to
that achieved by simple linear regression models, and only slightly superior to that of novices (Camerer & Johnson, 1991; Johnson, 1988). Most reports of this expert “mediocrity” have relied on domains in which there are no unequivocally correct rules but only sets of more or less accurate heuristics (Johnson, 1988), which human experts have difficulty applying and combining in the correct statistical manner. In consequence, performance in those domains can be optimized by forming weighted linear combinations of probabilistic cues, a process embodied in linear regression but apparently difficult to achieve by humans (Camerer & Johnson, 1991). To circumvent those difficulties, human experts use a variety of alternative combinatorial strategies. One of them, known as configural reasoning, consists of considering predictor variables in a categorical manner rather than by weighted addition. For example, a configural rule in medical diagnosis might be: “if the patient experiences headaches that have a gradual onset, with no periods of remission, and has high levels of spinal fluid protein, then diagnose a brain tumor” (Schwartz & Griffin, 1986, p. 94). Configural reasoning is often observed in experts, but unlike weighted linear regression its all-or-none character renders it vulnerable to small variability in measurements (Camerer & Johnson, 1991).

2.3 Adaptive Expertise

Thus far, we have limited our discussion to what some have described as “routine expertise”, in contrast to what is termed “adaptive expertise” (e.g., Gott, Hall, Pokorny, Dibble, & Glaser, 1993; Kimball & Holyoak, 2000). Adaptive expertise has been defined as “…an advanced level of problem-solving performance … characterized by principled representations of knowledge … as opposed to representations dominated by surface features” (Gott et al., 1993, p. 259).
Although routine and adaptive expertise are often seen as two contrasting concepts (e.g., Kimball & Holyoak, 2000), we are reluctant to accept this contrast for a variety of reasons. First, we are not aware of an independent criterion that identifies a particular expert, or a particular domain of expertise, as adaptive. Instead, expertise appears to be considered adaptive whenever it transfers well and it is considered routine whenever it does not. Second, empirical examinations of adaptive expertise converge on identification of the same, or similar, cognitive principles that are also involved in non-adaptive settings. For example, Barnett and Koslowski (2002) presented experienced restaurant managers and business consultants without any experience in the hospitality industry with problems relating to the management of hypothetical restaurants. Because the specific problems were novel to both groups of participants, Barnett and Koslowski considered them to represent “transfer” problems. Notwithstanding the lack of domain-specific expertise, the business consultants were found to outperform the restaurant managers, suggesting that the consultants were “adaptive” experts whereas the managers’ expertise was more “routine.” Further analysis identified the amount of prior consulting history (i.e., strategic business advisory experience) as the crucial variable underlying the performance difference. A crucial characteristic of business consulting, in turn, is the extreme breadth and variability of the problems that consultants tend to encounter. Barnett and Koslowski therefore conclude that “…a possible explanation for the observed differences is … the wide variety of business problem-solving experience to which the consultants, but not the restaurant managers, have been exposed” (p. 260).

As we review below, variability among training instances is a known strong predictor of transfer in general. We therefore propose that adaptive expertise does not
differ qualitatively from routine expertise, and that the observed differences in transfer 
ability are best explained within known principles of knowledge and expertise.

We now turn to an examination of those broader principles of knowledge in contexts 
other than expertise. This examination, in turn, also provides another, more theoretical 
perspective on the nature of expertise.

3. Structure of Knowledge

3.1 Overview

By discussing the structure of knowledge, we implicitly assume that knowledge can 
have a structure – that we can reasonably discuss constructs such as individual parcels of 
knowledge (what might once have been called “ideas”), pathways connecting the parcels, 
domains that the parcels apply to, dominance relations between parcels, and so on. As 
befits this general review, we abstain from detailed analysis of the structure of any 
particular domain and instead concentrate on a central dichotomy: is knowledge generally 
integrated or is it generally fragmented?

The idea that knowledge is highly integrated is a central, if often tacit, tenet of views 
of expert knowledge (e.g., Bédard & Chi, 1992; Ericsson & Lehmann, 1996). Glaser 
(1996), for example, explicitly cited the centrality of the “… acquisition of well-
organized and integrated knowledge that provides a structure for representation that goes 
beyond surface features. For the development of expertise, knowledge must be acquired 
in such a way that it is highly connected and articulated, so that inference and reasoning 
are enabled as is access to procedural actions” (p. 306). Finally, it has been assumed that 
during training, experts notice inconsistencies in their current knowledge, “… which in 
turn will serve as a stimulus for further analysis … until an acceptable reintegration … is
attained” (Ericsson, 1996, p. 38). To foreshadow our main point, in contrast to the prevailing views of expert knowledge, we will reach the conclusion that knowledge is often more fragmented and disconnected than experience would seem to indicate.

Discussions of knowledge integration are made difficult by the essentially intuitive nature of the term “integrated.” As diSessa and colleagues (diSessa, Gillespie & Esterly, 2004) point out, it is possible to analyze the term (they use “coherent” for our “integrated”) as consisting of a position on dimensions of contextuality, relational structure, and specification. Integrated knowledge (such as that described by Gott et al., 1993) is largely de-contextualized in that irrelevant surface features are not represented or considered; the color of a body in motion does not enter into the laws of thermodynamics. Integration also implies increased structure in the relations between elements, for example in that some elements are logically deduced from others. Integration also entails the ability of an observer to specify knowledge in a more compact manner. It takes less verbiage to describe the ideal and derived gas laws than it does to explain a naïve theory of balloons, drinking straws, sweating lemonade glasses and exploding soda pop cans.

The distinction between integration and fragmentation can also be expressed within a spatial metaphor; for example by noting the proximity of parcels to each other, or their size and relation to the size of the domain of interest. The spatial metaphor turns out to be at the heart of recent formal analyses of people’s knowledge.

Knowledge spaces. The notion that knowledge may be organized spatially relates to the acknowledgment that similarity can be understood as a spatial relationship. In other words, the ‘space’ in which one can represent knowledge elements is a space in which distances are defined by a metric of similarity. Shepard (1987) pointed out that the ability
to generalize afforded by a spatial representation of similarity is a potential solution to one of the core problems in psychology; how new knowledge can be related to old. This relationship is critical in order for people to learn as rapidly as they do, given the type of data they are presented with; as Landauer and Dumais (1997) say, this leads to a potential solution of “Plato’s problem”; namely, induction.

Landauer and Dumais (1997) introduced a method known as Latent Semantic Analysis that can derive and represent the relationships between knowledge parcels as spatial relations. Latent Semantic Analysis (LSA) assumes that words with similar meanings are likely to occur in the company of similar words; put another way, if words appear in similar contexts then they likely have similar meanings. LSA uses large corpora of text in which paragraph-length sections serve as context. Each word in the corpus is tallied for each context, and statistical dimension-reduction techniques are used to summarize the context-word relationships. Each word is then identified as a point in a space of lower dimensions than the number of contexts, but still of relatively high dimensionality (approximately 300 dimensions works well for many applications). The model assumes that all words can be represented in a single semantic space, and in this sense is clearly integrated rather than fragmented.

The hypothesis that a single semantic space is an effective way to represent general knowledge about word meanings is supported by Vigliocco, Vince, Lewis, and Garrett’s (2004) Featural and Unitary Semantic Space model (FUSS). FUSS differs from LSA in that its metric dimensions are interpretable; this is because the data for FUSS are not word-context relations in corpora but speakers’ judgments of the semantic features of words. In a related vein, Griffiths and Steyvers (2002) used corpora to arrive at
interpretible dimensions in their Topics model, although their statistical technique for dimension reduction sees similarity as a relation between probability distributions rather than spatial locations (a move similar to that taken by Probabilistic Latent Semantic Analysis, Hoffman, 1999).

None of these models, in their current forms, purport to be fully adequate descriptions of the sum of an individual’s knowledge about the meanings of words in their language. However, the move to represent meanings in integrated formats, within a spatial metaphor of some form, is one that must be recognized.

Alternative approaches. The theories of knowledge spaces just discussed stand in contrast to several alternative approaches that deserve mention. These alternative approaches have variously involved concepts such as feature lists (e.g., Barsalou & Hale, 1993), marker networks (e.g., Berg, 1992; Delisle, Moulin, & Copeck, 2003; Lange, 1992), schemata (e.g., Kintsch, 2000; Marshall, 1995), or mental models.

The concept of a mental model has a particularly long and venerable history in applied and industrial settings. Mental models are thought to consist of look-up tables, lists of formal declarative knowledge, or mental images that contain representations of objects, functions and causal interconnections between objects necessary to perform a task (see, e.g., Moray, 1999). Mental models offer a variety of functions such as providing rules for predicting future states from current information (Bellenkes, Wickens, & Kramer, 1997). For example, expert pilots expend more gaze time on predictive instruments than novices, indicating that experts utilize their mental model for predicting future states more than novices (Bellenkes et al., 1997).
3.3 Co-existence of alternative knowledge

We now examine evidence for the continued co-existence of alternative ways in
which people use their knowledge at all levels of skill acquisition and expertise (e.g.,
Lovett & Schunn, 1999; Shrager & Siegler, 1998). Reder and Ritter (1992) and Schunn,
Reder, Nhouyvanisvong, Richards, and Stroffolino (1997) presented participants
repeatedly with two-digit × two-digit multiplication problems (e.g., 43 × 19). Before
responding, participants had to rapidly indicate whether they could retrieve the correct
answer from memory (which they then had to report immediately) or whether they would
need to compute the answer (in which case extra time was allotted). Most relevant for
present purposes is the finding that across repeated presentations of a given problem,
people were found to switch strategies not just once but between 2 and 3 times, and
switches were separated by up to 50% of all learning trials (reported in Delaney, Reder,
Staszewski, & Ritter, 1998). This suggests that both forms of knowledge—retrieval and
computation—continued to co-exist throughout the experiment.

Prolonged co-existence of alternative knowledge has also been observed at a much
larger time scale, namely across grades in primary school (e.g., Shrager & Siegler, 1998;
Siegler, 1987). This research showed that children approach single-digit mental
arithmetic with immense cognitive variability, and that some techniques—such as
counting fingers vs. retrieving the answer from memory—may co-exist for several years
and may compete for selection whenever a problem is presented. Correspondingly, even
adult performance can be characterized by an interaction between memory retrieval and
alternative strategies (Griffiths & Kalish, 2001). Griffiths and Kalish showed that many,
but not all, systematic aspects of the pattern of errors observed in simple multiplication
problems (Campbell, 1994) could be explained by the similarity (and thus confusability) of the problems as predicted by a retrieval-based response strategy.

The persistence of competing knowledge structures is consonant with the suggestion that people ordinarily maintain multiple parcels of knowledge that could apply to any given situation (diSessa, 1988). This suggestion has two non-trivial implications: First, it presupposes that there is a selection process that can choose among plausible alternative parcels. This selection process is presumably based on the structural and perceptual similarities of the current situation with those stored in memory (Gentner, 1989). Second, if the ordinary state of knowledge includes multiple overlapping parcels, then at least some of those parcels might contain mutually inconsistent and contradictory information. This, indeed, appears to be the case.

3.4 Knowledge partitioning

Consider first an instance of contradictory knowledge that, though consolidated by experience, is at the lower end of the expertise spectrum. Tirosh and Tsamir (1996) reported inconsistencies in high school students’ understanding of the concept of mathematical infinity. Depending on the surface structure of the problem presentation, the distribution of responses differed greatly: Whereas with one surface structure, 80% of participants correctly identified two infinite sets as containing the same number of elements, the vast majority of the same respondents (70%) gave the opposing, inconsistent answer with the other surface structure. In a related study, also involving mathematical knowledge, Even (1998) showed that few prospective secondary mathematics teachers spontaneously linked an expression to its isomorphic graphical representation, even though this linkage would have facilitated solution of the problem.
The reverse was also true: people had difficulty deriving an expression from a graphical representation of the same function. Given that subjects were highly conversant with both representations of all functions used in the study, the findings by Even point towards heterogeneity even in consolidated knowledge.

Contradictory elements of knowledge have also been revealed in another naturalistic domain known as “street mathematics.” This research focused on people who lack formal schooling but are able to solve mathematical problems in every-day contexts, for example young street vendors, fishermen, construction foremen, and cooks in Brazil (e.g., Carraher, Carraher, & Schliemann, 1985; Nunes, Schliemann, & Carraher, 1993). Notwithstanding their minimal formal schooling, the participants in those studies were highly competent at solving mathematical problems associated with their domain of expertise.

Of greatest interest here is a context manipulation involving expert cooks (Schliemann & Carraher, 1993). Participants were presented with identical proportionality problems either in a pricing context (“If 2 kg of rice cost 5 cruzeiros, how much do you have to pay for 3 kg?”), or in a recipe context (“To make a cake with 2 cups of flour you need 5 spoonfuls of water; how many spoonfuls do you need for 3 cups of flour?”). Importantly, both problem contexts were familiar to participants and relevant to their domain of expertise. Schliemann and Carraher reasoned that social convention dictated accuracy in the pricing context, whereas estimation might be acceptable for recipes. Those expectations were confirmed: In the pricing context, subjects used a variety of identifiable mathematical strategies in preference to estimation, with the result...
that accuracy was in excess of 90%. In the recipe context, by contrast, accuracy was dramatically lower (20%) and half of the responses given were based on estimation.

In the preceding cases, contradictory performance arose between variants of problems that differed not only according to the context in which they were presented (e.g., their cover story) but also their surface structure. An even purer instance of contradiction, involving reasoning about materially identical problems that differed only in an irrelevant context, was observed in the study by Lewandowsky and Kirsner (2000) that was mentioned earlier. Lewandowsky and Kirsner (2000) asked experienced wild fire commanders to predict the spread of simulated wild fires. The experts’ predictions were found to depend on an additional variable, the physically irrelevant problem context. When a fire was presented as one that had to be brought under control, experts nearly always expected it to spread with the wind. When an identical fire was presented as a “back burn,” experts predicted the reverse, namely that the fire would spread uphill and into wind. Back burns are fires that are lit by fire fighters in the path of an advancing to-be-controlled fire to starve it of fuel; back burns obey the same laws of physics as any other fire, in the same way that apples and oranges both obey the laws of gravity.

Before presenting an explanatory framework for these results, it is essential to differentiate them from conventional context effects, such as those reviewed earlier that underscored the specificity of expertise. Four attributes of the Lewandowsky and Kirsner study are relevant in this regard: (1) The nature of the problem and its surface structure arguably did not differ between contexts. That is, unlike the conventional context effects in expertise, the problem was no more typical of the domain in one context than the other. (2) By implication, unlike the related study by Schliemann and Carraher (1993), the
change in context was a minimal alteration of a verbal label that accompanied presentation of a problem. (3) Both domain-relevant contexts were part of the training regime of the experts and both regularly occurred in the field. (4) The context shift resulted not only in a reduction of performance, as for example observed with chess masters’ memory of random board configurations, but in a qualitative reversal of the response. That is, the same problem yielded two mutually exclusive and contradictory predictions, each of which was consistent with application of a domain-relevant predictor variable.

These attributes are sufficiently unique to warrant the assertion that knowledge, even within a well-learned domain, may exhibit little homogeneity. Indeed, it appears that experts may sometimes, perhaps often, have knowledge that is overlapping and contradictory. As we observed earlier in this section, overlapping knowledge parcels are indicative of fragmented knowledge structures, and fragmentation has been assumed to be the norm for naïve theories (diSessa et al., 2004). Lewandowsky and Kirsner (2000) suggested that the fragmentation observed in experts be considered an example of knowledge partitioning, caused by associative learning and thus a natural consequence of acquiring expertise.

Lewandowsky, Kalish, and Ngang (2002) proposed that associative learning produces knowledge partitioning in essentially the following way. Early in learning when few cases are known, the learner acquires information rapidly about the few available cases. As learning continues, the most effective strategy to deal with new problems that are similar to the learned cases is to use the initially-learned information. Thus, it is effective for learners to protect their old knowledge and apply it whenever it is
applicable. People achieve this protection through rapid shifts in attention (Kalish, Lewandowsky, & Kruschke, 2004). This process of learning new cases when necessary and deflecting change from old cases creates knowledge parcels that may contain contradictory information. So long as the stored cases do not overlap, this is not a problem for the learner, and so the associative theory predicts that partitioning is only sustainable when such conflict does not routinely occur. In the firefighting example, this may indeed have been the case as wildfires tended (in the experts’ experience) to occur during high-wind periods and back-burns tended to be encountered (or set) primarily when winds were light.

Our discussion of knowledge “parcels” is not to give the impression that knowledge representations are necessarily static. On the contrary, there is evidence that knowledge, specifically conceptual knowledge, is not static and may be created or altered “on-the-fly”. For example, knowledge assembly theory (Hayes-Roth, 1977) posits that repeated activation of the same components leads to the unitization of these components into a configuration which is then activated as a single, integrated entity. As another example, Barsalou’s (1983) work with ad hoc categories has demonstrated that highlighting or making salient a particular goal can alter one’s judgement about an item’s category membership, its typicality, and the activation of other related items (Barsalou, 1982, 1983, 1985). More recently, Barsalou (1999) has taken the notion of “on the fly” recruitment of knowledge even further, by suggesting that knowledge is fundamentally linked to physical experience. According to his perceptual symbols theory, knowledge of a concept is represented as a modality-specific “simulation” of one’s physical experience with that concept. For example, the sweetness of a strawberry is represented by
“simulating” (or imagining) its taste. Knowledge is, thus, fragmented according to modalities, of which Barsalou identifies six: vision, audition, taste, smell, touch, and action (Pecher, Zeelenberg, & Barsalou, 2004). Which of these six is activated for simulation depends on the context in which the concept is encountered (Pecher et al., 2004), thus further underscoring the dynamic—and fractionated—nature of knowledge representations within this framework.

Given the ready occurrence of partitioning and fragmentation, the apparent integration of knowledge in the expert may now appear all the more remarkable. However, close examination of the way experts apply their knowledge suggests that this appearance is at least partially an illusion. We have suggested that knowledge is frequently accessible only from an associated context, or, equivalently, that knowledge is often represented at a grain size that is smaller than the domain the knowledge ought to (in a normative sense) cover. One measure of this grain-size is the ease of transfer; problems within a knowledge parcel’s domain should see transfer, the knowledge should be used with more difficulty on problems outside the parcel’s boundaries. In the next section, we take up this measure and evaluate the integration of knowledge with respect to transfer.

4. Transfer of Knowledge

The use of existing knowledge in new situations, known as transfer, is perhaps the most important test of one’s current knowledge structures. Transfer necessarily involves linking or mapping from what is known to a new or novel situation (Holland, Holyoak, Nisbett, & Thagard, 1986). This mapping entails a trade-off between expediency, which requires the rapid application of knowledge, and efficiency, which requires the selective
application of only those cognitive resources that are necessary for the task (Besnard & Cacitti, 2004). Inherent in this trade-off is the potential for transfer to fail.

Failure of transfer can have drastic consequences in an applied setting. Besnard and Cacitti (2005) described an industrial accident at a French steel factory, where the installation and use of new machinery amongst several older machines resulted in the death of a factory worker. The worker was operating a thread drawing machine, a device used to reduce the diameter of a metal thread by gradually increasing its tension. The output of this machine is wound tightly onto a drum and held in place by a set of pressing wheels controlled by the operator. On the new machine, the two key buttons controlling the opening and closing of the pressing wheels were swapped with respect to the older machines. The experienced operator mistakenly opened the pressing wheels on the new machine at a time when the metal thread was tightly wound, causing the thread to uncoil violently and resulting in the death of the worker. Prior experience with the old machines led the worker to transfer an existing skill to a situation that required similar skills, but applied in different manner, with deadly results.

This is not to conclude that successful transfer is impossible or rare; we have already seen that experts are extremely adept at transferring their knowledge within their domain of expertise. There is also evidence for successful within-domain (also called “near” transfer) among novices. For example, having learned a specific rule to categorize stimuli, people are able quickly to learn to categorize novel stimuli that share the same rule but are instantiated by different dimensions (Shepard, Hovland, & Jenkins, 1961).

However, as we show later, the extent of transfer between tasks often falls short of what intuition might lead one to expect. For example, in the domain of artificial grammar
learning, transfer is much better if the surface structure of the training set remains the same during the test phase (Brooks & Vokey, 1991; Gomez, Gerkin, & Schvaneveldt, 1994). Slight contextual changes (e.g., replacing colours with colour names; Dienes & Altmann, 1998) can reduce or eliminate transfer altogether.

We now examine the conditions that determine whether or not transfer is successful. We focus on cognitive factors and do not consider variables that are beyond the scope of this chapter, such as organizational factors (e.g., perceived support; Flint, 2003), characteristics of the individual (e.g., IQ, Ceci & Ruiz, 1993, Ceci, Rosenblum, & DeBruyn, 1999; motivation, Bereby-Meyer & Kaplan, 2005), or social factors like mentoring or supervisor support (e.g., Cromwell & Kolb, 2004).

4.1 Similarity and Transfer

For transfer to occur, people must necessarily perceive two tasks as being similar. The emphasis on perception is crucial, because transfer depends primarily on the perceiver’s psychological processing rather than objective measurements of the tasks involved. We consider four factors that are known to affect the perception of similarity.

Perceived similarity: Structure vs. surface. Perhaps the most important differentiation between forms of similarity involves the distinction between “deep” structural similarity and “surface” similarity, which comprise two potentially independent means of describing the relations between two situations or objects. For example, two fables that involve completely different sets of characters (and hence share little surface similarity) may nonetheless make the same moral point (thus having identical deep structure). Conversely, two fairy tales may involve the same set of characters but provide completely different messages. The latter situation can be particularly harmful because
when surface similarity lures people into attempting transfer between tasks that are structurally dissimilar, negative transfer may result (Hershey & Walsh, 2000). For example, a novice attempting to understand the game of cricket might mistakenly apply his or her knowledge of American baseball because a bat and a ball are used in both games. This attempt at transfer is fatal because the deep structure of cricket—which is sufficiently grave and complex to be summarized not by mere rules but by “laws”—deviates considerably from the comparatively simple deep structure of baseball.

Conversely, a change in cover story or surface presentation can reduce transfer notwithstanding deep structural identity between the two tasks. For example, in a now classic study, Gick and Holyoak (1980, 1983) taught participants to solve a problem involving the storming of a fortress surrounded by a mine field, in which the key to successful conquest was to send numerous platoons from all directions simultaneously that then converged onto the target. After learning this solution, people were unable to apply that knowledge to an isomorphic radiation convergence problem, in which removal of a tumor without damaging the surrounding tissue could only be achieved by applying weak intersecting radiation from all directions (Gick & Holyoak, 1980, 1983). Hence, while transfer to similar problems is indeed possible, a surprisingly small change in context or cover story can eliminate that transfer quite readily.

**Similarity and expertise.** The distinction between surface and structural similarity is particularly relevant when comparing the transfer abilities of novices and experts. One of the primary differences between expert and novice problem solving is that experts focus on the deep structure of the task (e.g., Chi et al., 1981; Dunbar, 1995). Accordingly, experts will attempt transfer if two tasks share structural similarities even if their surface
similarity is low. As we have already seen, mathematical expertise predicts the ease with which people transfer between superficially different word problems with the same structure (Novick, 1988). Novices, by contrast, will only attempt transfer if there are salient surface similarities between the source and the target, despite the fact that the same solution process is needed (Cormier, 1987). For example, in the earlier convergence problems, novices, after being trained in the radiation context, are more likely to attempt transfer to a problem that is similar at the surface, because it involves x-rays, rather than to a structurally similar problem that is less similar at the surface because it involves ultrasound (Holyoak & Koh, 1987).

**Conceptual vs. structural similarity.** Dixon and colleagues (Dixon, Zimmerman, & Neary, 1997; Dixon & Gabrys, 1991) differentiate between conceptual similarity, which is based on information about why a procedure works or how a device operates and allows the application of conceptually similar problem solving steps, and structural similarity, which is similarity based on the steps that must be performed to solve a problem and allows for the application of identical procedural steps. In their studies, participants were initially trained to operate a complex device through a series of sub-goals comprised of a number of different steps (e.g., for an airplane device, the sub-goal ‘Engine Start Up’ might consist of the steps ‘engine 1’, ‘engine 2’, followed by ‘ignition’). Consequently, conceptual similarity (which, in this example, is isomorphic to superficial similarity) could be manipulated independent of structural similarity by changing the order of the sub-goals but maintaining the same order of steps within each sub-goal.
Following initial training with one device, transfer to a second superficially unrelated device (e.g., an alarm system) was impaired compared to transfer to a superficially related device (e.g., an airplane with differently labeled controls) when the order of sub-goals was changed. Transfer was poorer still when the order of steps within the sub-goals was changed, compared to when the steps were unchanged, regardless of whether the sub-goal order was also altered. Hence, the order of the steps comprising the deep structure was not as important to maintaining acceptable transfer as the order of the routines within these steps (Dixon et al., 1997).

**Similarity and context.** The context in which a judgment is made can greatly alter perceived similarity. For example, changing the context during problem solving can affect encoding of the problem, which in turn can either facilitate or deter successful transfer. In a “pass-along” task, in which blocks of various sizes are shifted within a frame from an initial configuration to a known goal-state, completion of a difficult problem becomes easier if an analogy can be identified between the difficult problem and an easier one (Zamani & Richard, 2000). For instance, in the difficult problem, two rectangular blocks are encoded as either two halves of a square or as two separate rectangles depending on whether an easier problem with an identical solution, but with a square block instead of two rectangular blocks, is shown first or not at all (Zamani & Richard, 2000). Furthermore, for the difficult problem to be used as an analogue for an even harder problem, both problems must share the same goal state. Recognition of similar goal states allows for the application of the same solution strategies for both problems (Zamani & Richard, 2000). If the two difficult problems have different goal
states, and hence different solution procedures, knowledge is not transferred from one to the other (Zamani & Richard, 2000).

The context of training can also affect the judgment of similarity between problems. For example, in category learning, if people are trained to associate a context cue with a particular region of category space, then despite the fact that the context cue does not predict category membership, people will use context to gate their responses and will treat an identical stimulus differently in two different contexts (Yang & Lewandowsky, 2003, 2004; see also the earlier discussion of Lewandowsky & Kirsner, 2000).

4.2 Failures of Transfer

When people fail to perceive the similarity between what they know and a novel task, transfer does not occur. People may fail to note relevant isomorphisms for a variety of reasons.

Context specificity. Transfer fails more readily if the first task is more context-specific than the second one. For example, if people are initially trained to answer physics problems and are then tested with more general algebra problems, which nonetheless involve the same concepts, transfer is poor (Bassok & Holyoak, 1989). Conversely, if people are initially trained with algebra problems, transfer to physics problems remains intact (Bassock & Holyoak, 1989).

Extent of Learning. Failures of transfer can stem from failures of learning, for example when training involves only a limited number of problems (see, e.g., Catrambone & Holyoak, 1990; Loewenstein, Thompson, & Gentner, 1999). Likewise, if training involves only prototypical examples, then transfer will only be possible for target problems that are suitably similar to the prototypes (Elio & Anderson, 1984; see also
Gick & Holyoak, 1987). The inverse of this statement, that transfer is greater if training involves a broader range of problems, is also true and we have already noted that it may underlie apparent instances of “adaptive” expertise.

Accordingly, techniques that improve learning have also been shown to improve subsequent transfer (Aleven & Koedinger, 2002). For instance, compared to rote learning, transfer is better after learning that required participants to generate solutions to problems (Flint, 2003) or test hypotheses (Burns & Vollmeyer, 2002). Similarly, training which emphasizes different objectives can facilitate the transfer of different skills (Bitan & Karni, 2003). For example, when trained to recognize nonsense words (e.g., PON, LOP) composed of a Morse-code like script (e.g., PON = |^*|[^*]|^, LOP = *^||[^*]), people were able to transfer knowledge of the specific “letters” (i.e., they could recognize novel “words” composed of the letters used at training) only when initially instructed on how to decode the script. When people were trained on non-alphabetical words (i.e., the Morse-code like script did not consistently map to specific letters), they were unable to transfer any of the learning to novel words (Bitan & Karni, 2003). Importantly, people who learned how to decode the script performed much worse on old words comprised of new symbols than people trained on non-alphabetical words. The demonstration of both positive and negative transfer within the same condition illustrates the differential effects of training with different objectives.

Negative Transfer. Negative transfer is said to occur when performance on a novel task following training is poorer than it would have been without any prior training. Although typically not accompanied by the fatal consequences that struck the unfortunate French steel worker, negative transfer can occur whenever surface similarities mask
structural differences. For example, Woltz, Gardner, and Bell (2000) trained subjects in a complex number reduction task in which the rule for reduction of a larger number to a smaller number was determined by the relationship between the first two digits of the larger number (e.g., if the two digits differ by a value of two, replace those two digits with the midpoint between the two digits, or if two digits are equal remove the first digit). Participants were initially trained solely on stimuli that required the application of one sequence of rules. For example, the numbers 3565 and 9767 both require application of the “midpoint” rule for the first and second reduction and the “equal” rule for the final reduction (e.g., 3565 becomes 465, 465 becomes 55, and 55 becomes 5). Subjects produced errors when new sequences were presented at transfer that initially resembled training sequences but that required the application of a different rule for the last reduction (e.g., 3567 requires application of the midpoint rule twice to reduce the number to 57, but the final reduction is different to the training stimuli). Hence, for these “garden path” stimuli, the initial similarity in rule application masked the necessity of a different final rule and thus resulted in negative transfer compared to stimuli which were either completely dissimilarity or highly similar to the training stimuli.

Moreover, when task complexity was increased (e.g., by increasing the number of possible rules), increased training not only led to increased positive transfer for novel sequences (i.e., faster response latencies), but also led to an increase in undetected errors for new “garden path” sequences (Woltz et al., 2000, Experiment 2). In this case, as in the case of the French factory worker, increased training and expertise led to enhanced negative transfer.
These instances of negative transfer, which usually occur spontaneously, have often been referred to as “strong-but-wrong” errors (Norman, 1981). Reason (1990) linked strong-but-wrong errors to a process of frequency gambling and similarity matching. That is, if a process has been successful in the past, people are more likely to continue applying that same process if it appears similar to the target task (see also Lovett & Anderson, 1996). Negative transfer is distinct from other forms of error, such as simply computing an incorrect response from a correct algorithm, because the performance decrement involves the application of prior knowledge or training in a situation that does not require it. In Woltz et al.’s (2000) number-reduction task, we can distinguish between negative transfer and calculation error because errors on “garden path” problems were committed with the same speed as correct responses to training problems. By contrast, latencies for novel regular transfer sequences were longer than for training sequences. One implication of negative transfer is that the resultant “strong-but-wrong” errors go unnoticed and thus escape the possibility of discovery and correction (Woltz et al., 2000).

**Increasing Positive Transfer.** On the basis of the preceding discussion, one might be tempted to assume that simply informing people about the similarity between two tasks might enhance transfer. Contrary to that intuition, it turns out that transfer is facilitated if the process of discovering and extracting similarities between tasks, particularly similarities between relational information, is self-initiated (Dixon & Dohn, 2003). In their study, people were given problems consisting of different numbers of alternating, connected beams, each supported by a fulcrum, which acted in a see-saw fashion with the action of the first beam affecting the action of the second and so forth. People were told which direction the first beam was moving, and were asked to predict the direction of
movement of the final beam. Participants were either told to classify each beam as an *up* beam or a *down* beam in alternation or were given no instructions. Half of the people who did not receive instructions quickly discovered the up/down strategy, while all of the people given the up/down instructions used that strategy exclusively. When shown new problems involving gear systems, those people who received no instructions and nonetheless discovered the up/down strategy also quickly discovered an analogue of the up/down strategy and applied it to the gear problems. The people who received instructions, however, fell back to a less efficient tracing strategy. Hence, people demonstrated better transfer when allowed to discover more efficient strategies for themselves, a process labeled “redescription” by Dixon and Dohn (2003). It follows that training regimes which allow for self-discovery should lead to more effective transfer, although this notion has yet to be tested.

5. Conclusions

We have touched on a variety of issues in research on knowledge and expertise. At the risk of glib over-simplification, we propose to condense our review into the claim that knowledge is best understood by rejecting its existence as a coherent concept: Instead of talking about “knowledge,” we prefer to talk about a *set of highly context-specific learned responses*. Adopting this perspective appears to be particularly useful for the practitioner because it automatically accommodates the following major limitations and shortcomings of knowledge and expertise:

1. Expertise is highly domain-specific and brittle. Accordingly, expert performance can suffer dramatically if the deep structure of a task is altered, even if only slightly. In an applied setting, any alteration of an expert’s domain is likely to result in decreased
performance. This decrease is likely to be more severe if the deep structure of the task, such as the number and sequence of steps involved in performing a task or a task rule, is altered. It follows that practitioners should ensure that the steps and rules in the current task closely match the steps and rules in the expert’s domain of expertise.

(2) While expertise transfers well within a domain, little or no transfer can be expected outside a domain. In general, transfer often falls short of what intuition might lead one to expect because it occurs only if people correctly perceive two tasks to be similar. In practice, it is crucial that practitioners wanting to ensure positive transfer maximize the likelihood of two tasks being perceived similarly. A pertinent example of this is the release of software updates. If the updated software changes the settings so that different labels are given to identical function, the user’s performance will suffer, until he or she is well practiced with the new version.

(3) Negative transfer can result, with potentially grievous consequences, if people are misled by the surface similarity between two tasks with very different deep structures. The resulting “strong-but-wrong” errors often escape detection and correction. In practice, this means that tasks which require different operations should be given different surface features to minimize negative transfer. This is particularly true when two tasks with different operations are required in close temporal or spatial proximity.

(4) Knowledge frequently reveals itself to be fragmented and partitioned, with people exhibiting quite contradictory behaviour on an otherwise identical problem in different contexts. Even experts, who are typically assumed to have a highly integrated knowledge base, may exhibit surprisingly contradictory behaviour. In response,
practitioners should ensure that training occurs in a variety of contexts, and should make trainees aware of the potential hazards of partitioned knowledge.

(5) Experts also often exhibit expediency; that is, the tendency to master a task by focusing on a few key variables at the expense of ignoring other, and sometimes crucial information. One way to tackle this issue is to highlight the importance of considering all information during training. If training highlights the pitfalls of ignoring relevant information, perhaps by designing a task where ignoring relevant information leads to failure, then the learner will hopefully be more aware of this shortcoming of expertise. It has also been suggested that providing the opportunity for learning by self-discovery can result in more flexible knowledge that is readily transferred.

Although the preceding list of limitations and shortcomings is perhaps sobering, it need not detract from the stunning achievements that people are capable of. Whether considered “knowledge” or a “set of context-specific responses,” abilities such as the retention of 30,000 digits of \( \pi \) or 50,000 chess patterns, or the ability safely to operate a machine as complex as an Airbus A380, are remarkable by any criterion.
References


Knowledge and Expertise

56

Footnote

1 By the same token, research has identified domains in which exceptional performance cannot be detected. For example, people who claim to be speed readers have been found to exhibit remarkable dexterity at turning pages without displaying any comprehension of the text (Homa, 1983). Those “domains” are commonly excluded from consideration in research on expertise.
Table 1

Commonalities between experts in different domains identified by Holyoak (1991).

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Experts perform complex tasks in their domains much more accurately than do novices</td>
</tr>
<tr>
<td>2.</td>
<td>Experts solve problems in their domains with greater ease than do novices</td>
</tr>
<tr>
<td>3.</td>
<td>Expertise develops from knowledge initially acquired by weak methods</td>
</tr>
<tr>
<td>4.</td>
<td>Expertise is based on the automatic evocations of actions by conditions</td>
</tr>
<tr>
<td>5.</td>
<td><strong>Experts have superior memory for information related to their domains</strong></td>
</tr>
<tr>
<td>6.</td>
<td>Experts are better at perceiving patterns among task-related cues</td>
</tr>
<tr>
<td>7.</td>
<td>Expert problem-solvers search forward from given information rather than backward from goals</td>
</tr>
<tr>
<td>8.</td>
<td>One’s degree of expertise increases steadily with practice</td>
</tr>
<tr>
<td>9.</td>
<td>Learning requires specific goals and clear feedback</td>
</tr>
<tr>
<td>10.</td>
<td><strong>Expertise is highly domain-specific</strong></td>
</tr>
<tr>
<td>11.</td>
<td>Teaching expert rules results in expertise</td>
</tr>
<tr>
<td>12.</td>
<td>Performances of experts can be predicted accurately from knowledge of the rules they claim to use.</td>
</tr>
</tbody>
</table>
Table 2

Expert Shortcomings (items 1-8 were identified by Holyoak, 1991)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Inflexibility</td>
</tr>
<tr>
<td>2.</td>
<td>Expediency</td>
</tr>
<tr>
<td>3.</td>
<td>Mediocrity</td>
</tr>
<tr>
<td>4.</td>
<td>Inefficiency</td>
</tr>
<tr>
<td>5.</td>
<td>Poorer memory for cases outside domain</td>
</tr>
<tr>
<td>6.</td>
<td>Poorer perception of patterns unrelated to expert performance</td>
</tr>
<tr>
<td>7.</td>
<td>Asymptotic performance</td>
</tr>
<tr>
<td>8.</td>
<td>Domain specificity</td>
</tr>
<tr>
<td>9.</td>
<td>Subjectivity</td>
</tr>
<tr>
<td>10.</td>
<td>Lack of knowledge integration</td>
</tr>
<tr>
<td>11.</td>
<td>Variation in strategies</td>
</tr>
<tr>
<td>12.</td>
<td>Knowledge Inaccessibility</td>
</tr>
</tbody>
</table>
Chapter 10

Error discounting in probabilistic category learning

Error Discounting in Probabilistic Category Learning

Stewart Craig, Stephan Lewandowsky, and Daniel R. Little

University of Western Australia

Running Head: Error Discounting

Address correspondence to: Stewart Craig

School of Psychology (M304)
University of Western Australia
Crawley, W.A. 6009, AUSTRALIA
+61 8 6488 7862 (voice)
craigs03@student.uwa.edu.au
URL: http://www.cogsciwa.com/
Abstract

Some current theories of probabilistic categorization assume that people gradually attenuate their learning in response to unavoidable error. However, existing evidence for this error discounting is sparse and open to alternative interpretations. The presence of error discounting was investigated in three experiments by training participants to categorize single-dimensional cues on the basis of probabilistic reinforcement that was shifted at some point during training. In all three experiments, participants displayed evidence of slowed learning indicative of error discounting. Quantitative modeling of the data revealed that adding a mechanism to handle error discounting significantly improved the fits of both an exemplar-based and a rule-based associative learning model.

Supplementary material relevant to this article can be downloaded from http://www.cogsciwa.com. (Web material will be moved to http://www.psychonomic.org/archive upon acceptance of this article.)
Throughout daily life, people make decisions based on uncertain cues that are predictive of multiple outcomes: Doctors make diagnoses based on symptoms that are indicative of several diseases; we learn to judge whether or not a dark cloud will bring rain, or which sports team is likely to win on the weekend. Performance in such situations requires learning of the probabilistic relationship between the predictive cues and the uncertain outcomes (Kruschke & Johansen, 1999).

Probabilistic categorization is one form of probability learning in which the to-be-predicted outcomes are discrete, with distinct categorical values rather than continuous magnitudes. Unlike its deterministic counterpart, in probabilistic categorization the same stimuli belong to multiple categories albeit with differing probabilities. In consequence, the same response to an identical stimulus can be reinforced as either correct or incorrect on a probabilistic basis (Kruschke & Johansen, 1999). This corrective feedback is typically believed to drive category learning (Gluck & Bower, 1988a, 1988b; Rumelhart, Hinton, & Williams, 1986).

In conventional category learning, people often perform sufficiently well to eliminate error completely. In probabilistic categorization, by contrast, avoiding errors and achieving perfect accuracy is by definition impossible. On that basis, Kruschke and Johansen (1999) proposed that during probabilistic categorization, people may eventually begin to accept a certain level of error as unavoidable and may thus progressively slow their learning, a phenomenon known as error discounting. The idea of error discounting is supported by the finding that information introduced midway through training is utilised less than equivalent information present from the outset (Edgell, 1983; Edgell &
The notion of error discounting is central to at least one view of probabilistic category learning (Kruschke & Johansen, 1999).

This article critically examines the notion of error discounting and seeks supporting evidence for its operation in probabilistic categorization. We proceed as follows: We first examine existing evidence for discounting and find it to be surprisingly sparse. We next present three probabilistic categorization experiments in which the reinforcement probabilities for all stimuli were shifted during training, with the abruptness and timing of the shift varying between experiments. People’s responses were found to track those shifts with a notable delay, as expected on the basis of error discounting. Finally, we explore alternative mechanisms that may underpin error discounting by applying three models of categorization to our results: An exemplar model without an associative learning mechanism (GCM; Nosofsky, 1986), another exemplar model based on associative learning (RASHNL; Kruschke & Johansen, 1999), and a rule-based model that also included associative learning (derived from the hybrid ATRIUM model; Erickson & Kruschke, 1998, 2002). We find that only the associative-learning models can handle the data at all, and that both provide a significantly better account if they incorporate error discounting. We conclude that error discounting deserves to be given a prominent theoretical role in probabilistic categorization.

Error Discounting in Probabilistic Environments

Initial evidence for error discounting consisted of the finding that delayed introduction of cues reduces their utilization. If cues suddenly become relevant after an initial period of non-relevance, people use them less than equivalent cues that were relevant from the outset (Edgell, 1983; Edgell & Morrissey, 1987). This finding is
compatible with the notion that people discount error later in learning. In consequence, participants fail to notice a change in the environment and fail to learn about a highly relevant cue (Kruschke & Johansen, 1999). Although error discounting is compatible with the data, other explanations for the findings by Edgell and colleagues cannot be ruled out: For example, the observed non-utilization of cues after their delayed introduction may be the result of conventional blocking (Kamin, 1969; Rescorla & Wagner, 1972).

Blocking occurs when a previously learned association between a cue, $A$, and an outcome, $X$, prevents learning of a new association between a second cue, $B$, and $X$, when trained in the presence of $A$ (Kruschke & Blair, 2000). For example, if one has learned to associate a noise with an outcome (e.g., electric shock), then after further training with the conjunction of (the same) noise and a (novel) light, the light on its own will show little if any evidence of learning on a final test. By analogy, it is possible that the newly-relevant cue in the Edgell studies was under-utilized because it was blocked by the association between the already-predictive cue and the outcome.

In order to rule out a blocking explanation, it is preferable to seek evidence for error discounting in situations in which there is only one stimulus dimension present: with a single cue, blocking cannot occur. The single-cue approach recognizes that if error discounting is present, people should become less sensitive to shifts in the reinforcement probabilities associated with the training stimuli. For example, if a stimulus is initially placed in category $A$ with 80% probability, and this suddenly shifts to 20% at some point during learning, error discounting implies that people might continue to predominantly assign that stimulus to category $A$, at least for some time.
Relevance Shifts in Probabilistic Environments

Research on the effects of shifts in reinforcement probabilities dates back several decades (Estes, 1984; Estes & Straughn, 1954; M. P. Friedman et al., 1964; Yelen & Yelen, 1969). However, upon closer inspection, those data turn out to be either ambiguous or of little relevance to the error discounting notion.

In a series of early studies (e.g., Estes & Straughn, 1954; M. P. Friedman et al., 1964; Yelen & Yelen, 1969) participants learned to predict which of two lights would be illuminated on the next trial. The actual probabilities with which the lights were illuminated differed from chance, and by associating the actual outcome (identity of the illuminated light) with their prediction, people’s responses gradually came to track the actual event probabilities. Most relevant here is the fact that in all studies those probabilities changed during the experiment. For example, in the study by Yelen and Yelen (1969), one light might be illuminated on 90% of the first 100 trials (and the other light during the remaining 10%), with those probabilities suddenly reversing for the next 100 trials. In virtually all cases, participants were found to adapt to those changes with remarkable speed. At first glance, this outcome might seem to compromise the notion of error discounting. Upon closer inspection, however, the data arguably have little bearing on probabilistic categorization processes because people never associated distinct outcomes to different stimuli: The only “stimulus” in those early studies was the signal that denoted the beginning of a trial, and consequently any observed learning did not necessarily involve the formation of associations between stimuli and outcomes but simple awareness of base-rates of reinforcement. The data of Estes and Straughn (1954), M. P. Friedman et al (1964), and Yelen and Yelen (1969) thus confirm that people remain
sensitive to base-rates even after more than 1,000 trials of training with varying probabilities (M. P. Friedman et al., 1964); however, they tell us little about error discounting in probabilistic categorization.

More recently, in a closer analogue to probabilistic categorization, Estes (1984) asked participants to choose between two alternatives on each trial. Each alternative, in turn, was associated with two outcomes (success or failure) that were reinforced with some varying probability. For one of the alternatives, the probability of success remained constant at .5 throughout. For the other alternative, the probability of success increased and decreased throughout training following a sine pattern. The results were intriguing: On the one hand, Estes found that participants were generally aware of the changing probability structure for one of the outcomes and responded accordingly (i.e., following the cyclical sine pattern, preferring that alternative whenever its probability of success exceeded .5 and rejecting it whenever it fell below .5). On the other hand, during a final transfer block involving a uniform probability of success of .5 for both alternatives, people persisted with the cyclical pattern and showed little evidence of adapting to the new reinforcement structure.

On balance, the existing data involving relevance shifts are, at best, ambivalent with respect to error discounting. The study that most closely resembled probabilistic categorization (Estes, 1984), simultaneously found evidence both for people’s ability to track changing probabilities and against their ability to adapt after prolonged training. The latter outcome is suggestive of error discounting but must remain inconclusive.
Current Study

The current experiments built on these precedents by examining more systematically how people adapt to changes in a probabilistic categorization task. The experiments used a single, quasi-continuous cue that could take on four values and was probabilistically associated with two categories. This task has two key benefits: first, previous studies have found that people can readily learn to categorize such probability structures (Jones, Love, & Maddox, 2006; Little & Lewandowsky, in press) and second, the fact that only a single cue is present precludes the possibility of blocking.

The principal manipulation was instantiated by switching the reinforcement probabilities of the two categories part-way through training. All experiments shared a common method, only varying the rate—sudden or gradual—and timing—early or late in training—of the shift in reinforcement. A gradual shift may limit the increase in error, thus perhaps causing a greater tendency to discount error than if the shift is sudden. Similarly, the tendency to discount error may increase with the extent of learning, suggesting that a later shift might give rise to more discounting.

To facilitate exposition, we name the three experiments after the type of shift and the number of training blocks presented before the shift: Thus, we refer to Experiments 1 to 3 as Sudden-10, Gradual-10, and Gradual-6, respectively. We first report the data from all experiments before seeking a quantitative account within alternative models.

General Method

Participants

Forty-four members of the authors’ campus community participated, with 24 participants completing Experiment 1 (Sudden-10); 10 completing Experiment 2
(Gradual-10); and a further 10 completing Experiment 3 (Gradual-6). Participants either received course credit or a remuneration of $10.

Stimuli and Apparatus

The experiment was controlled by a Windows computer using a MATLAB program created with the aid of the Psychophysics toolbox (Brainard, 1997; Pelli, 1991). The task involved a single-cue binary choice that required participants to classify a series of squares of varying sizes into one of two categories, \( A \) or \( B \). The training stimuli were four squares of linearly increasing sizes of 44, 66, 88, and 110 mm edge lengths. An MDS solution obtained for squares of this type has shown that participants perceive the change in size approximately linearly (Colreavy & Lewandowsky, in press). The linearity of perception simplifies data presentation and modelling.

Squares were presented in black on a white background. Actual category membership of each stimulus was determined on a probabilistic basis, with the probability of an item belonging to category \( A \) at the outset of training equalling .2, .4, .6, and .8, respectively, in increasing order of size. Category \( B \) probabilities were \( 1 - P(A) \). Part-way through the experiment, those probabilities were shifted as explained below.

Procedure

Each trial commenced with a fixation symbol (“+”) in the centre of the screen for 500 ms. The “+” was replaced by the stimulus square, which was shown in the centre of the screen along with the response options (i.e., Category F or Category J) underneath, both of which remained on screen until a response was made by pressing the ‘F’ or ‘J’ keys. Mappings of the two response keys to categories \( A \) or \( B \) were randomly determined for each participant. After each response, feedback (CORRECT or WRONG)
was presented on the screen, below the item, for 1,300 ms. Incorrect responses were additionally followed by a beep.

Items were randomly presented in blocks of 40 trials, with each square presented 10 times per block. In all experiments, participants were trained for 18 blocks altogether. Participants received self-timed breaks, lasting a minimum of 30 seconds, after every four blocks. The session lasted approximately 45 minutes.

Experiment 1. In Experiment 1 (Sudden-10), participants were trained for 10 blocks before the probabilities shifted, after which training continued for a further 8 blocks. The shift occurred instantly, with the probability of an item belonging to category A becoming .8, .6, .4, and .2—and correspondingly, .2, .4, .6, and .8 for category B—from Block 11 onward.

To facilitate learning, participants were informed of their accuracy, given by percentage correct, for the preceding block after each of the first 8 blocks. (Note that this implies that accuracy feedback ceased before the reinforcement probabilities shifted.)

An additional manipulation in Experiment 1 involved the attempt to induce participants in one condition to respond by “maximizing”; that is, by responding consistently with A (or B) whenever P(A) > .5 (or P(B)>.5). Because this manipulation had no bearing on the error-discounting results, it is reported in the Appendix.

Experiment 2. In Experiment 2 (Gradual-10), people were likewise trained for 10 blocks before the probabilities shifted, but the change was gradual, with probabilities of items being in category A first changing to .35, .45, .55, and .65 (Blocks 11 and 12), then to .65, .55, .45 and, .35 (Blocks 13 and 14), before reaching their final assignment of .8, .6, .4, and .2 (Block 15 onward). As before, the probabilities that an item belonged to
category B shifted in unison with the category A probabilities, so that \( P(B) = 1 - P(A) \).

Feedback was again provided during the first 8 blocks only.

**Experiment 3.** In Experiment 3 (Gradual-6), participants were trained on the original probabilities for 6 blocks (the first 4 of which were followed by feedback). The probabilities then shifted gradually, first changing to .35, .45, .55, and .65 (Blocks 7 and 8), then to .65, .55, .45 and, .35 (Blocks 9 and 10), before reaching their final assignment of .8, .6, .4, and .2 (Block 11 onward).

**Results and Discussion**

**Learning Measure**

To gauge the extent of learning, probability-matching (PM) scores were calculated in all experiments for the participants’ mean responses during the last two pre-shift blocks, based on the following formula:

\[
PM_j = (R_i(A \mid j) - P(A \mid j)) \times SI_j, \tag{1}
\]

where \( P(A \mid j) \) is the training probability for item \( j \), and \( R_i(A \mid j) \) is the probability of participant \( i \) responding with category A in response to item \( j \). \( SI_j \) is a sign indicator for item \( j \), such that if \( P(A \mid j) < .5 \), then \( SI_j = -1 \), else if \( P(A \mid j) > .5 \) then \( SI_j = 1 \) (D. Friedman & Massaro, 1998). Each participant’s PM score was based on the average across the four items.

For ease of interpretation, the PM scores were then transformed to a scale ranging from 0 (chance responding) to 1 (full maximizing). For the present probability structure, a
score of .4 indicates perfect probability matching. (Scores below 0 indicate reversed responding.)

Experiment 1 (Sudden-10)

Three participants had PM scores that were indicative of chance responding (0.1, 0.02, and -0.1). Because these participants either did not understand the task or made no attempt to learn it, they were excluded from further analysis. The mean PM score of the remaining participants ranged from 0.3 to 0.93 with a mean of 0.58 (SD = 0.19). A t-test showed the overall mean PM score to be significantly above chance, $t(20) = 14.07$, $p < .001$, confirming that people had learned the category structure prior to the shift.

Mean response probabilities for the four training items across all blocks are presented in Figure 1. The figure shows that responses diverged quickly from chance in Block 1. Learning then continued to progress, albeit at a slowed rate, with only a gradual increase in mean response probability from the end of Block 1 to the end of the pre-shift blocks. After the shift, by approximately Block 12, mean response probabilities for all items had transited through .5, demonstrating that participants adapted to the shift in probabilities. Notably, the rate at which learning progressed after the shift was slower than during initial learning. Participants took noticeably longer to differentiate between the stimuli after crossing through .5 than at the start of training.

In order to permit a more concise summary of people’s adaptation to the shift in training probabilities, responses within each block were collapsed into a single slope. By design, $P(A \mid j)$ across $j$ is linear within a block: Given that participants largely probability-matched, as evidenced by the mean PM score, the slope of the line through $R_i(A \mid j)$ across $j$ within each subject-block was likely also linear (we noted earlier that
the stimuli are known to be perceived linearly), a possibility confirmed by inspection of responses. Regression analyses were thus conducted to determine the slope of the linear trend in response probabilities for each block for each participant. Mean response slopes and the slopes of the objective training probability for all blocks are presented in Figure 2. The figure shows that response slopes initially showed a positive trend which gradually increased to a level above the slope of the training probabilities. After the shift, the slopes declined, crossing through zero between Blocks 11 and 12 to become increasingly negative. The slower post-shift learning was also evident, because whereas the confidence intervals of the response slope bracketed the actual slope from Block 1 onward (until an overshoot emerged around Block 7), in the post-shift environment the confidence intervals did not bracket the training slope for 4 blocks of training (and an undershoot did not begin to emerge until the final block). This finding suggests that people exhibited some level of error discounting.

Inspection of individual responding suggested that the aggregate pattern was reflective of behaviour at the individual level. An indication of individual responding is given by the confidence intervals in Figure 2.1

Experiment 2 (Gradual-10)

PM scores were again calculated for participants for the last two pre-shift blocks. These ranged from 0.3 to 0.88 ($M = 0.60, SD = 0.20$) and all participants were retained for the analysis. A $t$-test showed that the PM scores were significantly greater than chance, $t(9) = 9.38, p < .001$.

Figure 3 shows the mean response probabilities. Overall, participants adapted to the shift. There was, however, notably little change in responding after the first shift in
training probabilities (i.e., in Blocks 11 and 12). Participants appeared to adapt more to the second shift—that is, in Blocks 13 and 14—with all response proportions transiting through .5 by Block 15.

The corresponding response slopes are shown in Figure 4. Again, there was little change in slope after the initial shift (i.e., after Block 10), and even after the second shift (i.e., after Block 12) slopes changed only after a further block of training (i.e., in Block 14). Thus, participants demonstrated little adaptation to two consecutive shifts in reinforcement probability for approximately 3 blocks, or 120 trials. This delay in adaptation to the shift strongly suggests the presence of error discounting.

From Block 14 onward, the reduction in slope was more dramatic and people ultimately completed the probability shift by the end of the experiment. As in Experiment 1, participants’ learning appeared slower after the shift than during initial learning, as indicated by the fact that the confidence intervals of the slopes did not bracket the true values for 5 blocks of training after the shift commenced.

**Experiment 3 (Gradual-6)**

Two participants were excluded from the analysis because their PM scores in the last two pre-shift blocks (-0.13 and -0.15) were below chance. PM scores for the remaining participants ranged from 0.3 to 0.68 (\( M = 0.50, SD = 0.13 \)). A t-test showed the mean PM score to be significantly above chance, \( t(7) = 11.18, p < .001 \).

Figure 5 shows the response probabilities for each item and Figure 6 shows the underlying slopes. Response patterns were generally consistent with the previous experiments. As in Experiment 2 (Gradual-10), participants learned quickly at the outset but showed little change during the initial shift. The observed slopes transited through
zero around Block 11. Reduced learning was again evident after the shift, and by the end of training, the observed mean response slopes had not reached the objective values.

There was seemingly less overall adaptation to the final probabilistic environment in Experiment 3 than in Experiments 1 and 2, in which the response slopes had crossed through the objective slopes by the end of training. The greater adaptation with more initial training is consistent with a finding known as the overtraining reversal effect, in which people who are given more initial training appear more willing to adapt to a shift in the environment (Juola & Hergenhahn, 1967; Kruschke, 1996). However, at first glance, the greater adaptation observed with more training runs counter to the notion of error discounting; we therefore now model our data to examine whether they are compatible with error discounting.

Computational Modeling

We applied three models to the data: the GCM (Generalised Context Model; Nosofsky, 1986), RASHNL (Rapid Attention SHifts 'N' Learning; Kruschke & Johansen, 1999), and ATRIUM (Attention to Rules and Instances in a Unified Model; Erickson & Kruschke, 1998). These models were chosen based on the following considerations: We first examined the GCM, which does not contain an associative learning mechanism, to consider an alternative explanation for our results based on sample size. To foreshadow the outcome, the application of the GCM ruled out that alternative explanation.

We next examined RASHNL, an exemplar model in which associative learning plays a prominent role. Particularly relevant here is the “annealing” mechanism in RASHNL, which gradually reduces the model’s learning rates to instantiate error
Error Discounting

16
discounting. The modelling revealed that the data were accommodated significantly better when annealing was present.

Finally, we examined a rule-based variant of ATRIUM, which also relies on associative learning, to extend the generality of our conclusion regarding annealing. We again found that the data were better accommodated when annealing was present.

We conclude that annealing-based error discounting, irrespective of whether it is embodied in an exemplar-based or rule-based model, plays an important and identifiable role in probabilistic categorization.

GCM: A Sample-Size Explanation

It is possible that the decline in post-shift learning revealed by the present experiments could be an automatic by-product of the inevitable increase in item sample size during learning. Specifically, in the very early stages of training, when few items have been presented, on an exemplar view each further individual stimulus will have a relatively large impact. However, at later stages of learning, new items become increasingly insignificant in relation to the overall number of exemplars already encountered, thus limiting their impact. The delayed adaptation to the shift could therefore result from the small impact of items later in training, relative to items early in training. We explored this alternative within the GCM (Nosofsky, 1986). The GCM contains no associative learning mechanism but represents all encountered instances in memory, thus providing a quantitative instantiation of the sample-size hypothesis.
GCM Specification. The GCM assumes that on each trial the current item activates all previously encountered stimuli stored in memory according to:

\[ s_{ij} = \exp(-c \cdot d_{ij}), \]

where the similarity, \( s_{ij} \), between items \( i \) and \( j \) is determined by the distance, \( d_{ij} = |x_i - x_j| \), which is the psychological distance between items \( i \) and \( j \) in (one-dimensional) psychological space (Nosofsky, 1986). (Note that for simplicity of exposition, all equations reported in this article are tailored to the fact that our stimuli were unidimensional.) The specificity parameter, \( c \), determines the sharpness of the exponential function.

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

\[
P(A_1i) = \frac{\left( \sum_{j \neq i} s_{ij} \right)^\gamma}{\left( \sum_{j \neq i} s_{ij} \right)^\gamma + \left( \sum_{j \neq i} s_{ij} \right)^\gamma},
\]

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

GCM Simulations. The GCM was fit separately to each participant’s mean response probabilities for all items in all blocks across all experiments. The four stimuli were
Error Discounting

18

coded as the integers 1 to 4. The GCM was presented with the training sequences shown to participants. Parameters were estimated using the SIMPLEX algorithm (Nelder & Mead, 1965) to minimise the negative binomial log-likelihood:

$$-\ln L = -\sum_i d_i \ln(p_i) + (n_i - d_i) \ln(1 - p_i),$$  \hspace{1cm} (3)

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

Table 1 shows the GCM’s estimated parameters for all three experiments. As shown in Figure 7, the GCM failed to capture the data in some crucial ways: In particular, it was unable to readjust its predictions in response to the probability shift. This failure is readily explained by considering the large number of exemplars presented before the shift. Even in Experiment 3 (Gradual-6), each item was presented 60 times prior to the shift, and because the model’s predictions were based on the average responses across all presentations, the GCM can never accumulate enough new evidence within the number of training trials to reverse its predicted probabilities. (Note, however, that there is a downward trend in the predicted slopes after the shifts, which indicates that given massively extended training, the GCM’s performance might come to mirror the final outcome probabilities). We conclude that our results cannot be accommodated by an explanation based solely on sample size.\(^2\)
RASHNL: Error Discounting via Annealing of Learning

Formal Description of RASHNL. RASHNL (Kruschke & Johansen, 1999) is an exemplar-based connectionist model of probabilistic categorization that was developed as an extension of ALCOVE (Attention Learning COVering map; Kruschke, 1992), which is itself an extension of the GCM (Generalized Context Model; Nosofsky, 1986). Central to RASHNL is the concept of annealing of learning rates. This annealing mechanism captures error discounting by gradually decreasing the rate at which the model learns over time.

The annealed learning provides the model with various advantages over a fixed learning rate: Basic research on annealed learning in neural networks has shown that it allows for large changes in responding early in training, to quickly adapt to the environment, while the slowing of learning later in training permits the fine-tuning of probabilistic responding (Amari, 1967; Bös & Amari, 1998; Heskes & Kappen, 1991; Murata, Kawanabe, Ziehe, Müller, & Amari, 2002). Annealed learning may also help RASHNL avoid unduly high sensitivity to order among stimuli late in training, such as observed in ALCOVE (Lewandowsky, 1995). The limited tests of RASHNL available to date have consistently found that the annealing-based error discounting mechanism improves the ability of the model to account for both probabilistic (Kruschke & Johansen, 1999) and non-probabilistic (Blair & Homa, 2005) categorization behaviour.

RASHNL has a layer of input nodes that correspond to the dimensions of the stimulus. In the present case, the activation of the single input node, \( d \), was equal to the psychological scale value, \( \psi \), of the presented stimulus.
As RASHNL is an exemplar-based model, new items are categorised based on their similarity to previously encountered category members (Kruschke & Johansen, 1999; Nosofsky, 1986). Hence, the input nodes connect to a layer of hidden exemplar nodes, which correspond to the training stimuli. Activation of the $j$th exemplar node is given by:

\[ h_j = \exp(-c|\psi_j - d|), \]  

where $c$, the specificity, is a free parameter that determines the slope of the gradient of the ‘receptive field’ of each exemplar; that is, the slope of the exponential decline in similarity with increasing distance between the current stimulus and the $j$th stored exemplar. (The present stimuli were uni-dimensional; accordingly, all equations are simplified to reflect this situation. As a further consequence of the unidimensionality, we removed RASHNL’s gain activation, attention shifting, and attention updating mechanisms, all of which only apply to multi-dimensional stimuli.)

Exemplar nodes connect to output nodes, which correspond to the available categories. Activation of the $k$th category node, $a_k$, is given by:

\[ a_k = \sum_j w_{kj} h_j, \]  

where $w_{kj}$ is a weight associating each exemplar with each category. Category activations are mapped onto response probabilities using a version of Luce’s (1963) choice rule, such that the probability of categorising a stimulus into category $K$ is determined by the
exponentiated activation of category $K$ over the sum of the exponentiated activation of all categories, given by:

$$\Pr(K) = \frac{\exp(\varphi \cdot a_K)}{\sum_k \exp(\varphi \cdot a_k)}, \quad (6)$$

where $\varphi$ is a scaling parameter representing decisiveness. If $\varphi$ is large, then a small activation advantage for category $K$ will result in large preference for category $K$, corresponding to maximizing behaviour. Conversely, if $\varphi$ is small, the response will be more uncertain and in proportion to the relative activations, thus corresponding to probability matching.

RASHNL is an error-driven learning model, such that each response is followed by feedback indicating the correct category in the form of teacher values for each category node. The error associated with a response is given by:

$$E = \frac{1}{2} \sum_k (t_k - a_k)^2, \quad (7)$$

where $t$ is the teacher value, such that $t_k = 1$ if the stimulus is a member of category $k$, and $t_k = 0$ if the stimulus is not a member of category $k$. 
Error Discounting

Learning proceeds through the minimization of $E$ via adjustment of association weights by gradient descent on error, given by:

$$\Delta w_{ij} = \lambda (t_k - a_k) h_j,$$

where $\lambda$ represents the learning rate. Crucially for the current experiments, this learning rate is annealed, such that as training progresses the rate of learning is slowed. Although several annealing mechanisms have been explored within the neural network literature (Amari, 1967; Bös & Amari, 1998; Heskes & Kappen, 1991; Müller, Ziehe, Murata, & Amari, 1998; Murata et al., 2002), RASHNL uses a ‘search and converge’ mechanism to decrease learning rates (see e.g., Darken & Moody, 1991). On each trial, $t$, the initial learning rate is multiplied by an annealing factor, $r$, given by the equation:

$$r(t) = 1/ (1 + \rho \times t),$$

where $\rho$ is a freely estimated, non-negative, scheduling constant that controls the rate of the annealing process (Kruschke & Johansen, 1999). Larger values of $\rho$ lead to faster annealing, and when $\rho$ is clamped at zero, the model exhibits no annealing at all (and saves a parameter in the process). This annealing function allows the model to make large shifts in learning early in training; from around trial $1/\rho$ onward, the learning rates rapidly reduce and converge to zero. In addition to the annealing rate, $\rho$, this version of
RASHNL had three other free parameters: specificity, $c$, the probability-mapping parameter, $\varphi$, and the weight-learning rate, $\lambda$.

**RASHNL Simulations.** The model was fit to the data in the same manner as the GCM, with the additional constraint that the output association learning rate was capped at 3.0. We compared two versions of RASHNL: One in which the annealing parameter, $\rho$, was freely estimated and another one in which it was set to zero.

Models were compared by AIC (Akaike Information Criterion; Akaike, 1974; Wagenmakers & Farrell, 2004), which adjusts for the number of parameters in a given model. The corrected AIC ($AIC_c$) is given by:

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

(10)

where $L$ is the maximum likelihood for the given model with $V$ free parameters taken over $n$ observations. The corrected AIC is recommended for use of samples where the ratio of data points to parameters is less than 40. The AIC thus combines two sources of information: Lack of fit (represented by the log likelihood) and a penalty term for model complexity (represented by the second and third terms in the above equation). $AIC_c$ values were further converted into AIC weights (Wagenmakers & Farrell, 2004). The AIC weights, $w_i(AIC)$, represent the conditional probability that the model $M_i$ is the best of the set of models tested.

Figure 8 shows the mean predictions of RASHNL with annealing turned on, and the $AIC_c$ values and AIC weights of the two versions of RASHNL (i.e., with and without
annealing) are shown in Table 3. For comparison, the table also provides the relevant statistics for the earlier fit of the GCM. Quite in contrast to the GCM, RASHNL captured the fast initial learning and moderately slowed post-shift learning that was displayed by participants. Mean (and median) best-fitting parameter values (aggregating across the fits to individual subjects) are shown in Table 2.

RASHNL was able to capture these features of the data because the annealing rates were relatively small compared to the learning rate, and considerably smaller than previously published values (see Table 2). This allowed the model to learn at a sufficiently high rate to adapt to the change in the environment; nonetheless, the estimated value of the annealing parameter was greater than zero, and the AIC values in Table 3 suggest that setting the parameter to zero impaired the fit of RASHNL.

To pin-point the role of annealing, we explored whether the loss of fit associated with setting $\rho$ to zero was statistically significant, using a likelihood-ratio test (Lamberts, 1997):

$$\chi^2 = -2[\ln L(\text{restricted}) - \ln L(\text{general})],$$

where $\ln L(\text{general})$ is the log-likelihood of the standard version of the model including all parameters (i.e., $\rho >0$), whereas $\ln L(\text{restricted})$ is the log-likelihood of the restricted version of the model, with annealing set to zero. Based on the likelihood-ratio test, RASHNL fit the data with annealing on significantly better than with annealing off, $\chi^2 (39) = 219.14, p < .001$. 
We conclude that annealing is crucial to enable RASHNL to capture people’s behavior in probabilistic categorization involving shifts in reinforcement. We next examine whether this role of annealing was tied to the particular architecture of RASHNL or whether it might represent a more general role that also applies to other models; for example, those based on rules. Rule models, such as general recognition theory (GRT; Ashby & Townsend, 1986) postulate that people categorise items based on an item’s distance from some rule boundary, and they have had some success in accounting for probabilistic categorization (Juslin, Olsson, & Olsson, 2003; Little & Lewandowsky, in press; Rouder & Ratcliff, 2004).

**rATRIUM: Annealing Without Exemplars**

The majority of rule-based models, including the GRT, do not include associative learning mechanisms. As an associative learning mechanism appears essential for investigating error discounting, we selected the rule module of ATRIUM (Erickson & Kruschke, 1998) as an alternative candidate model for the present data. ATRIUM’s rule module learns to associate rules with particular categories via a standard network learning algorithm, permitting implementation of annealing in the same manner as in RASHNL.

ATRIUM is a hybrid model that relies on both exemplars and rules; here, we eliminated the exemplar module because we were exclusively interested in the generality of annealing and its applicability within a rule-based architecture (hence the label rATRIUM for this variant of the model). rATRIUM divides the category space by a rule boundary set perpendicular to the relevant stimulus dimension. The stimulus dimension is represented by two rule nodes, \( r_{small} \) and \( r_{large} \), whose activations are given by:
\[ r_{\text{small}} = 1 - \frac{1}{1 + \exp[-\mu(d + \beta)]} \]  
(12)

and by:

\[ r_{\text{large}} = \frac{1}{1 + \exp[-\mu(d + \beta)]}. \]  
(13)

where \( d \) represents the value of a given item on the stimulus dimension. Each of these rule nodes forms a sigmoid threshold function, centered on the rule boundary, such that larger dimensional inputs will result in higher activation of the large rule node, while smaller dimensional inputs will result in a higher activation of the small rule node. The parameter \( \mu \) represents the gain of the sigmoid (i.e., its steepness), and thus controls the level of perceptual noise (or its equivalent) as dimensional values approach the rule boundary. Large values of \( \mu \) result in stimuli close to the rule boundary being more confusable. The parameter, \( \beta \), controls the position of the rule boundary.

The rule nodes are connected to output nodes that correspond to the possible category selections. The activation of output nodes, \( a_k \), for each category, \( k \), is calculated as the sum of the activations of the small and large rule nodes, given by:

\[ a_k = w_{k,\text{large}} r_{\text{large}} + w_{k,\text{small}} r_{\text{small}}. \]  
(14)

where the activation is moderated by the learned association weights, \( w_{k,i} \), between the rule and output nodes. As in RASHNL, the association weights are updated by minimising mean square error during learning. Thus, the association weights for each category, \( k \), and node, \( i \), are updated as follows:
where $\lambda$ is a freely estimated parameter which controls the rate of learning. Finally, output activations are converted into probabilities as in Equation 6 in RASHNL.

For present purposes, the annealing mechanism from RASHNL, given in Equation 9, was implemented in rATRIUM: Thus, an annealing rate, $\rho$, controlled the rate at which the weight learning rate, $\lambda$, was adjusted on successive trials. In summary, the model has four free parameters: A gain constant, $\mu$, which sets the standard deviation of the perceptual noise; a scaling constant, $\varphi$, that maps output probabilities to participant responses; the annealing rate, $\rho$; and learning rate, $\lambda$.

As is evident from the $w_j(AIC_c)$ values in Table 3, the fits of rATRIUM were better than those of the GCM, although they fell short of the fits of RASHNL (see Figure 9 for the mean predictions of rATRIUM). Nonetheless, like RASHNL, the rule model was able to adjust to the switch and capture the important aspects of the data. As with RASHNL, the fits of rATRIUM produced a small annealing rate, which allowed the model to adapt to the change in probabilities. Again like RASHNL, rATRIUM fit better with annealing on than when it was turned off, $\chi^2 (39) = 351.77, p < .001$.

The significant improvement associated with annealing suggests that irrespective of a model’s specific architecture, people’s adaptations to relevance shifts in probabilistic categorization are best modelled by including a mechanism for error discounting. To buttress this conclusion, we additionally examined whether individual differences between participants were related to the estimated annealing rates within RASHNL and
For each participant, across all three experiments, we computed the speed of adaptation to the probability shift by taking the number of blocks required for that person’s response slopes to shift by 0.4 (i.e., the actual range in slopes covered by the experimental manipulation). In a regression analysis, the estimated individual annealing rates were significantly and substantially related to speed of adaptation for both RASHNL, $R^2 = 0.23$, $F(1, 37) = 11.36$, $p < .01$, and rATRIUM, $R^2 = 0.16$, $F(1, 37) = 6.89$, $p < .05$. The longer people took to respond to the shift, the larger was their annealing rate estimate.

Summary of Modeling

The modeling suggests the following conclusions. First, a sample-size explanation without an associative learning mechanism is inadequate to account for the observed error discounting. Second, although annealing rates were generally small, two competing associative learning models both fit the data significantly better when their learning rates gradually declined across trials, implying that annealing is an essential component of models of probabilistic categorization. In further support, individual variation among participants in adaptation speed was captured by differences in the estimates of the annealing parameter within RASHNL. Third, the exemplar approach embodied in RASHL provided a better fit to the present data than a rule-based approach as instantiated in ATRIUM. Because emphasis here is on the general role of error-discounting rather than specific model comparison, we do not emphasize the latter result further.
General Discussion

Summary of Findings

The objectives of the current study were to investigate (a) whether people eventually discount errors during probabilistic categorization, and (b) what underlying mechanism might best explain error discounting, if present. Taken together, the results clearly suggest the presence of error discounting during probabilistic category learning; nonetheless, the discounting was far from complete as shown by people’s continued ability to adapt to shifts in the probabilistic environments. The slowed learning was primarily evident from (a) the difference between the initial fast learning and slowed post-shift learning, and (b) the lack of adaptation to the initial partial shift in Experiments 2 and 3. The present evidence for slowed learning was generally consistent with previous findings (Edgell, 1983; Edgell & Morrissey, 1987) that constituted a primary justification for error discounting (Kruschke & Johansen, 1999). Unlike those precedents, however, the uni-dimensional design of the present studies precluded alternative interpretations on the basis of blocking.

At a theoretical level, we showed that the slowed learning cannot be explained as a by-product of the increasing item sample size. Instead, the observed error discounting is best explained by an annealing of learning rates, as evidenced by the fact that both associative-learning models, RASHL and rATRIUM, fit the data significantly better when annealing was present than when it was turned off. Moreover, for both models, the extent of annealing was related to the extent of observed discounting at the level of individual participants.
The current results thus present clear, and novel, evidence that people slow their rate of learning and thus gradually discount error during probabilistic categorization. Before we discuss the implications of our results, we briefly take up some possible limitations.

**Limitations and Concerns**

Participants in our studies were not informed about a possible change in the probabilistic environment. In the natural world, by contrast, people may be more wary of changing probabilities and might thus adapt faster than observed here. However, the fact that learning was still slow after participants had transited through chance responding (i.e., a response slope of zero) suggests that being forewarned of a shift might not prevent error discounting.

At a theoretical level, we note that the use of a single stimulus dimension—while allowing the detection of error discounting in the absence of a blocking explanation—rendered the attention and gain features of RASHNL and rATRIUM irrelevant. As it is assumed in RASHNL that the attention and gain learning rates are also annealed (Kruschke & Johansen, 1999), the role of these features continues to await exploration.

At an empirical level, it likewise remains unclear how discounting functions in multi-dimensional environments. On the one hand, error discounting may play a greater role in complex environments, because if people have more difficulty determining the underlying probability structure, they may show a greater tendency to ignore errors. On the other hand, it may be that with multi-dimensional stimuli, error discounting is overshadowed by blocking effects, and thus has little impact on its own.
Exemplar vs. Rule Models

At present, it is unclear whether rule or exemplar models provide the preferable account of probabilistic categorization: Prior research has produced conflicting evidence, with either class of models finding favor with some researchers (Juslin et al., 2003; Kalish & Kruschke, 1997; McKinley & Nosofsky, 1995; Rouder & Ratcliff, 2004). It is quite likely that whether a rule- or exemplar-based strategy is used may vary depending on the conditions, stimuli, and individuals involved (Little & Lewandowsky, in press; Rouder & Ratcliff, 2004).

Although RASHNL provided a better fit to the data, the rule-based module from ATRIUM could also provide a respectable account. Thus, the present study does not rule out the possibility that some rule-based model may capture probabilistic categorization performance. In particular, there is no reason why annealing of learning rates should be exclusive to exemplar models.

Implications for Annealing

The modeling established that annealing of learning provided an advantage within both RASHNL and rATRIUM over parallel models without this feature. It must be noted, however, that the magnitude of annealing was quite small, particularly when compared to the previous parameter settings used by Kruschke and Johansen (1999). Thus, rather than causing a relatively dramatic decline in learning speeds, the present annealing levels imply a very gradual slowing of learning across training.

In addition, it must be noted that the evidence for error discounting was obtained using experiments in which reinforcement probabilities shifted during training. One might therefore question the generality of discounting, and in particular whether it is
ever-present, even in situations in which no shift occurs. This question is difficult to resolve empirically, for the reasons noted at the outset (i.e., the effects of delayed introduction of cues might be attributable to blocking). At a theoretical level, however, the answer is clear: When fitting RASHNL and rATRIUM to the pre-shift training blocks alone (details available in the web supplement), annealing rates were considerably greater, suggesting that annealing is required to handle all aspects of probabilistic categorization, rather than just circumscribed situations involving relevance shifts.
References


Different Response Behaviors

When performing probabilistic categorization, people display two primary response behaviors, known as probability-matching and maximizing, respectively (Fantino & Esfandiari, 2002; Shanks, Tunney, & McCarthy, 2002). Probability matching is said to occur if people place items into categories with a frequency approximately equivalent to the probability of reinforcement: For example, if stimulus $j$ is reinforced as belonging to category $A$ on 70% of all trials, then people probability match if they respond “$A$” with probability .7. By contrast, maximizing is said to occur if people always place an item into the category to which it is most likely to belong (Fantino & Esfandiari, 2002). Thus, for the above item $j$, maximizing implies that instead of responding with “$A$” on 70% of the trials, people would respond with $A$ on all trials.

Despite the fact that maximizing achieves the best attainable accuracy, people pervasively probability match (Fantino & Esfandiari, 2002; Shanks et al., 2002; West & Stanovich, 2003). Previous studies have found that rewarding participants and giving periodic accuracy feedback can increase the incidence of maximizing (Fantino & Esfandiari, 2002; Shanks et al., 2002). Similarly, people who tend towards probability-matching may be less likely to recognise probability concepts, such as independence of trials (Gal, 1996; West & Stanovich, 2003). Thus, teaching participants about such probability concepts, and providing performance-based rewards and accuracy feedback may help induce maximizing behaviour. This manipulation was included in Experiment 1 to permit exploration of any relationship between response strategies and error discounting.
Response Behavior Manipulation

The participants in Experiment 1 (Sudden-10) were randomly assigned to one of two conditions: 12 of the participants were placed into the maximising condition, and 12 into the probability-matching condition. In order to encourage maximising behaviour, participants in the maximising condition received 5c for each correct answer and lost 5c for each incorrect answer, rounded to the nearest dollar. Participants in the maximizing condition were also presented with the following question at the outset, based on that in (Gal, 1996), to demonstrate the advantage of maximising:

Suppose I have a die with four sides coloured red, and two sides coloured green. When rolled, if you win every time you correctly guess which colour lands up, on what proportion of trials should you select red to maximise your winnings?

In order to maximise, the correct answer is to select red on every trial. If participants did not answer correctly, they were informed of the correct answer and the reasoning behind this. The experimental trials did not commence until participants understood why the maximising response, in the given question, was the most effective strategy.

Response Behaviour Results

Mean probability-matching scores for participants in the maximising condition ($M = 0.56, SD = 0.16$) were slightly lower than mean probability-matching scores for participants in the probability-matching condition ($M = 0.60, SD = 0.22$), although the difference was not significant, $t(19) = 0.51, p > .05$. The inability to induce maximizing in participants is not entirely unexpected, because although the present techniques have been found to increase performance in some cases (Shanks et al., 2002), there are other instances in which people could not be encouraged to maximize (Fantino & Esfandiari,
Thus, data from all participants was aggregated across conditions for the remainder of the analysis and the manipulation was not considered further.
Author Note

We wish to thank Brian Ross and three anonymous reviewers for their helpful comments on previous versions of the manuscript. We also thank John Kruschke for providing the original RASHNL programming code. Preparation of this article was facilitated by a Discovery Grant from the Australian Research Council to the second author, and an Australian Professorial Fellowship to the second author.

Address correspondence to the second author at the School of Psychology, University of Western Australia, Crawley, W.A. 6009, Australia. Electronic mail may be sent to lewan@psy.uwa.edu.au. Personal web page: http://www.cogsciwa.com.
Footnotes

1 For a further examination of individual responding, see the web supplement for box-plots of participants’ transition of response slopes across blocks.

2 One obvious modification of the sample-size hypothesis involves the idea that memory representations of the presented exemplars decay over time, thus limiting the functional sample size. We explored several instantiations of that possibility in a series of additional simulations (reported in detail in the web supplement). None of those variants of the sample-size hypothesis alter our basic conclusions.
Table 1

Median (Mdn), mean (M), and standard deviation (SD) of estimated parameter values across participants (specificity, \(c\), and response scaling, \(\gamma\)) and negative log-likelihood values (-lnL) for the GCM fist to the Sudden-10 (S10), Gradual-10 (G10), and Gradual-6 (G6) experiments.

<table>
<thead>
<tr>
<th>Model</th>
<th>Exp.</th>
<th>Parameters</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>- lnL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mdn</td>
<td>M</td>
<td>SD</td>
<td>Mdn</td>
<td>M</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>GCM</td>
<td></td>
<td>(c)</td>
<td></td>
<td></td>
<td></td>
<td>(\gamma)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S10</td>
<td></td>
<td>2.87</td>
<td>6.26</td>
<td>8.50</td>
<td>1.12</td>
<td>1.37</td>
<td>0.85</td>
<td></td>
<td>9390.5</td>
</tr>
<tr>
<td>G10</td>
<td></td>
<td>1.38</td>
<td>5.93</td>
<td>10.07</td>
<td>1.32</td>
<td>2.38</td>
<td>2.12</td>
<td></td>
<td>4285.6</td>
</tr>
<tr>
<td>G6</td>
<td></td>
<td>1.50</td>
<td>1.76</td>
<td>1.27</td>
<td>1.33</td>
<td>4.04</td>
<td>7.48</td>
<td></td>
<td>3581.3</td>
</tr>
</tbody>
</table>
Table 2

Median (Mdn), mean (M), and standard deviation (SD) of estimated parameter values across participants and negative log-likelihood values (-lnL), for fits of RASHNL and rATRIUM to the Sudden-10 (S10), Gradual-10 (G10), and Gradual-6 (G6) experiments.

<table>
<thead>
<tr>
<th>Model</th>
<th>Exp.</th>
<th>Parameters</th>
<th>(\rho)</th>
<th>SD</th>
<th>Mdn</th>
<th>(\phi)</th>
<th>SD</th>
<th>Mdn</th>
<th>(\lambda)</th>
<th>SD</th>
<th>Mdn</th>
<th>(\gamma)</th>
<th>SD</th>
<th>(-\ln L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RASHNL</td>
<td>S10</td>
<td>0.003</td>
<td>0.130</td>
<td>3.55</td>
<td>0.18</td>
<td>0.40</td>
<td>0.53</td>
<td>1.22</td>
<td>12.00</td>
<td>34.91</td>
<td>8229.26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G10</td>
<td>0.007</td>
<td>0.006</td>
<td>3.16</td>
<td>0.25</td>
<td>0.21</td>
<td>0.16</td>
<td>1.30</td>
<td>4.56</td>
<td>9.88</td>
<td>3805.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G6</td>
<td>0.142</td>
<td>0.181</td>
<td>2.77</td>
<td>1.21</td>
<td>1.33</td>
<td>1.44</td>
<td>0.62</td>
<td>9.11</td>
<td>15.75</td>
<td>3415.41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S10</td>
<td>0</td>
<td>0</td>
<td>3.67</td>
<td>0.06</td>
<td>0.10</td>
<td>0.15</td>
<td>0.94</td>
<td>7.02</td>
<td>12.93</td>
<td>8273.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G10</td>
<td>0</td>
<td>0</td>
<td>3.23</td>
<td>0.04</td>
<td>0.07</td>
<td>0.07</td>
<td>1.32</td>
<td>4.28</td>
<td>9.65</td>
<td>3825.77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G6</td>
<td>0</td>
<td>0</td>
<td>3.12</td>
<td>0.08</td>
<td>0.27</td>
<td>0.54</td>
<td>0.39</td>
<td>1.11</td>
<td>1.19</td>
<td>3459.68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rATRIUM</td>
<td>S10</td>
<td>0.002</td>
<td>0.061</td>
<td>3.61</td>
<td>0.76</td>
<td>0.98</td>
<td>1.04</td>
<td>0.87</td>
<td>0.94</td>
<td>0.35</td>
<td>8286.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G10</td>
<td>0.009</td>
<td>0.015</td>
<td>3.28</td>
<td>1.33</td>
<td>1.17</td>
<td>1.03</td>
<td>0.91</td>
<td>0.98</td>
<td>0.38</td>
<td>3887.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G6</td>
<td>0.020</td>
<td>0.030</td>
<td>3.26</td>
<td>0.92</td>
<td>1.14</td>
<td>0.75</td>
<td>0.84</td>
<td>1.12</td>
<td>0.85</td>
<td>3342.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S10</td>
<td>0</td>
<td>0</td>
<td>3.67</td>
<td>0.18</td>
<td>0.35</td>
<td>0.35</td>
<td>0.87</td>
<td>0.84</td>
<td>0.34</td>
<td>8370.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G10</td>
<td>0</td>
<td>0</td>
<td>3.24</td>
<td>0.18</td>
<td>0.30</td>
<td>0.28</td>
<td>0.85</td>
<td>0.94</td>
<td>0.43</td>
<td>3922.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G6</td>
<td>0</td>
<td>0</td>
<td>3.27</td>
<td>0.24</td>
<td>0.27</td>
<td>0.24</td>
<td>0.80</td>
<td>1.04</td>
<td>0.88</td>
<td>3398.56</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3

Negative log-likelihood (-ln L), corrected Akaike Information Criterion values (AICc), and AIC weights, w(AIC) for fits of the GCM, RASHNL and rATRIUM across all experiments and participants.

<table>
<thead>
<tr>
<th>Model</th>
<th>Annealing</th>
<th>-lnL</th>
<th>AICc</th>
<th>w(AIC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCM</td>
<td>—</td>
<td>17257.40</td>
<td>34675.32</td>
<td>0</td>
</tr>
<tr>
<td>RASHNL</td>
<td>On</td>
<td>15449.82</td>
<td>31230.12</td>
<td>1</td>
</tr>
<tr>
<td>RASHNL</td>
<td>Off</td>
<td>15559.39</td>
<td>31363.05</td>
<td>0</td>
</tr>
<tr>
<td>rATRIUM</td>
<td>On</td>
<td>15516.16</td>
<td>31362.80</td>
<td>0</td>
</tr>
<tr>
<td>rATRIUM</td>
<td>Off</td>
<td>15692.05</td>
<td>31628.36</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: Bold entries indicate best model fit.
Figure Captions

Figure 1. Mean response probabilities for each of the four items across blocks in Experiment 1 (Sudden-10). The vertical line represents the point at which training probabilities were shifted. The legend provides the values of P(A) before and after the shift for each stimulus.

Figure 2. Slopes of the regression lines through the response probabilities across items within each block in Experiment 1 (Sudden-10). Error bars indicate 95% confidence intervals for participants’ response slopes. The dashed line represents slopes through the objective training probabilities.

Figure 3. Mean response probabilities for each of the four items across blocks in Experiment 2 (Gradual-10). Vertical lines indicate shift points for training probabilities. The legend provides the values of P(A) before and after the shift for each stimulus.

Figure 4. Slopes of the lines through the response probabilities across items within each block in Experiment 2 (Gradual-10). Error bars indicate 95% confidence intervals for participants’ response slopes. The dashed line represents slopes through the objective training probabilities.

Figure 5. Mean response probabilities for each of the four items across blocks in Experiment 3 (Gradual-6). Vertical lines represent shift points for training probabilities. The legend provides the values of P(A) before and after the shift for each stimulus.

Figure 6. Slopes of the lines through the response probabilities across items within each block in Experiment 3 (Gradual-6). Error bars indicate 95%
confidence intervals for participants’ response slopes. The dashed line represents slopes through the objective training probabilities.

Figure 7  GCM fits to Experiment 1 (Sudden-10); Experiment 2 (Gradual-10); and Experiment 3 (Gradual 6).

Figure 8.  Predicted slopes across item responses within each block when RASHNL was fit with annealing on. Error bars indicate 95% confidence intervals for participants’ response slopes. Dashed line represents slopes through objective training probabilities.

Figure 9.  Predicted slopes across item responses within each block when rATRIUM was fit with annealing on. Error bars indicate 95% confidence intervals for participants’ response slopes. Dashed line represents slopes through objective training probabilities.
Figure 1.
Figure 2.
Figure 3.
Figure 4.

Error Discounting
Figure 5.
Figure 6.
Figure 7.
Figure 8.
Experiment 1

Experiment 2

Experiment 3

Figure 9.
Chapter 11

Summary and Conclusions

The empirical and computational results of Chapters 6 and 7 offer several contributions.
1) Probabilistic feedback increased sensitivity to correlated cues. Probabilistic feedback led to a diffusing of attention across stimulus dimensions, which increased sensitivity to non-diagnostic category structure. 2) Chapter 7 demonstrated that knowledge partitioning can emerge in a probabilistic task, further demonstrating the ubiquity of the phenomenon. Selective attention was unable to account for this result. 3) Chapter 8 provided an account of how selective attention could be used to explain many of correlational sensitivity results in concept learning and concept usage tasks. 4) Chapter 9 provided a way of utilizing the knowledge partitioning framework in applied settings. 5) Finally, the results of Chapter 10 provided an empirical demonstration that learning in probabilistic environments is accompanied by an attenuation of learning.

One intriguing result from the empirical studies presented in this thesis was the finding that a simple measure of fluid intelligence predicted the emergence of knowledge partitioning. Further consideration of this result is presented below.

11.1 Knowledge Partitioning and Complexity

The correlation between knowledge partitioning and fluid intelligence is intriguing as it suggests that knowledge partitioning can be predicted by stable individual differences.
Chapter 7 suggested that partitioning might provide a way to simplify a complex problem. However, a quantitative analysis of knowledge partitioning is needed to determine whether partitioning does indeed reduce the complexity of a difficult problem. To provide such an analysis, we computed the complexity of knowledge partitioning representations and context insensitive representations of the discrete, Boolean dimensioned category space used in Chapter 6, the continuous dimension category space used in Chapter 7, and the continuous dimension space used in Yang and Lewandowsky (2004) and Lewandowsky et al. (2006). To compute representational complexity we followed the same method used to compute concept complexity (see e.g., Fass & Feldman, 2002; Feldman, 2006) but substituted response proportions for the category feedback which defined the concept space; see Chapter 12 for details of these computations.

The theoretical representations for each of these category spaces are shown in Figure 11.1. Panel A shows the category space from Chapter 6. The subpanel labelled data represents the category A stimuli from the training region of this category space; Category B stimuli are not highlighted in the panel as the analysis treats everything that is not a member of one category as a member of the opposing category. Panel B shows the category space from Chapter 7. Here again, the panel marked data represents the training region from those experiments (note that the gray region was not shown during training). The black areas represent category A. In the final panel, the category space used in Yang and Lewandowsky (2004) and Lewandowsky et al. (2006) is shown. In all three panels, the context insensitive (CI) subpanel shows the theoretical category representation used by participants who do not use the non-relevant context dimension to divide the category space into separate components. The knowledge partitioning (KP) subpanels show the theoretical representations used by participants who partition the category space on the basis of context.

Table 11.1 shows the relative complexity of each of the representations (see Appendix for details). In Chapter 7 and Yang and Lewandowsky (2004) the complexity of the KP representation is greater than that of the CI representation, but only if the KP
Table 11.1: Complexity of the category representations shown in Figure 11.1.

<table>
<thead>
<tr>
<th>Group</th>
<th>Complexity</th>
<th>Complexity Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 6: Little &amp; Lewandowsky (2008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CI</td>
<td>4</td>
<td>Algebraic Complexity</td>
</tr>
<tr>
<td>KP-Integrated</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>KP-Partitioned</td>
<td>2 + 2</td>
<td></td>
</tr>
<tr>
<td>Chapter 7: Little &amp; Lewandowsky (in press)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CI</td>
<td>13.18</td>
<td>Minimum Description Length</td>
</tr>
<tr>
<td>KP-Integrated</td>
<td>22.95</td>
<td></td>
</tr>
<tr>
<td>KP-Partitioned</td>
<td>11.48 + 11.48</td>
<td></td>
</tr>
<tr>
<td>CI</td>
<td>82816</td>
<td>Minimum Description Length</td>
</tr>
<tr>
<td>KP-Integrated</td>
<td>82990</td>
<td></td>
</tr>
<tr>
<td>KP-Partitioned</td>
<td>41495 + 41495</td>
<td></td>
</tr>
</tbody>
</table>

*CI = Context Insensitive; KP = Knowledge Partitioning.

The algebraic complexity measure is taken from Feldman (2006). The minimum description length measure is based on Fass & Feldman (2002). See the appendix for further details.

representation is considered to be an integrated whole (hence, the label KP-integrated). If instead, the KP group is assumed to divide the space into two components gated by context, then the complexity of each component on its own is less than that of the CI group. Yang and Lewandowsky (2003) provided evidence of such an effect by demonstrating that for the KP condition there was no difference in performance within each context from a separate condition which only learned one of the contexts. Hence, knowledge partitioning breaks the problem into two simpler components. Furthermore, this analysis of complexity can also explain why complete correlational sensitivity was not found in Chapter 6. As shown, in Table 11.1 both the CI and KP representations are of the same complexity; hence, there is no benefit gained by breaking the space into smaller components. Although each individual partition would be of less complexity, the complexity of the integrated KP representation is no more complex than the CI representation. The simplicity of the task environment has previously been found to be a boundary condition on the emergence of knowledge partitioning; that is, knowledge partitioning does not occur when the task is exceedingly simple (Lewandowsky et al., 2006). Another boundary condition appears when
we consider the limitations of types of correlations that people can learn. For example, people can not learn arbitrary cue correlations; instead they appear to be sensitive only to a limited number of simple relationships such as the one-to-one correlation between variables (Fass, 2006).

Although the complexity analysis can explain the correlation between knowledge partitioning and intelligence, it is unclear whether there is a general relationship between intelligence and selective attention. One possibility that is commensurate with the above analysis is that people with lower mental capacity have greater difficulty disregarding irrelevant information. Though this remains an avenue of inquiry to be further explored, developmental trends to attend to color cues over shape cues (Suchman & Trabasso, 1966) and a tendency to attend to compound cues over component cues at younger ages support this idea (Zeaman, 1978).

11.2 Conclusion

Probabilistic feedback increases attention shifting because attention shifts in response to error, and error is increased in probabilistic environments. Hence, because the world is comprised of a myriad of probabilistically related cues and outcomes, we are constantly seeking the most optimal or rational way to make decisions and solve problems. Through this increased attention shifting we acquire more knowledge about the world than is necessary. That is, we learn something about the non-relevant correlations that make our categories coherent, and add to the rich structure of the world. This correlational knowledge has immediate influence on the types of mental representations that we form and if used to partition our knowledge, enables the simplification of complex problems.

(b) Chapter 7: Little & Lewandowsky (in press)

(c) Yang & Lewandowsky (2004)

Figure 11.1: Data and theoretical context insensitive (CI) and knowledge partitioning (KP) hypotheses (i.e., representations).
References


Björkman, M. (1965). Learning of linear functions: Comparison between a positive


Human Performance, 3, 47-61.


Yang, L.-X., & Lewandowsky, S. (2004). Knowledge partitioning in categorization:


Chapter 12

Appendix

12.1 Complexity of Boolean concepts

There are several methods available for calculating the complexity of a Boolean concept, such as the concepts used in Chapter 6. Feldman’s (2006) concept algebra is a straightforward method based on the decomposition of a concept into three components: a) ”constant” properties, b) ”causally irrelevant” properties, and c) ”causally interacting” properties (Feldman, 2006, p. 346). Only causally interacting properties determine the complexity of a concept. Causally interacting properties are defined by considering the constraints that the values of one property places on the values of another property. For instance, consider a simple two dimensional concept comprised of discrete, Boolean dimensions X and Y. First, assume that the only positive members of the concept (or, equivalently, the only members of category A) occur when the X and Y dimensions take on the values 00, 01, or 11, respectively, and, by implication, the XY pair 10 is not a member of the concept. For this concept, X is constrained by Y because it can not take on a value of 1 unless Y has also taken on a value of 1. Note that Y’s value is not constrained by X and can therefore take on any value, regardless of X’s value. These constraints can occur at multiple hierarchical levels as the number of stimulus dimensions increases (i.e., X could be constrained by Y and Z in a three-dimensional concept or Y, Z, and C in the four-dimensional concepts used in Chapter 6). The complexity of a concept is computed by counting the number of
constraints of different degrees (i.e., hierarchical levels) and weighting the counts by the
degree of the constraint (see Feldman, 2006 for further details).

12.2 Complexity of concepts with continuous dimensions

The concept algebra outlined above does not provide any extension to continuously val-
ued stimulus dimensions. However, for the concepts used in Chapter 7 and Yang and
Lewandowsky (2004), minimum description length (MDL) techniques can be employed to
compute the relative complexities of the knowledge partitioning and context insensitive
representations. For the concepts used in Chapter 7, we use a variant of the method
described in Fass and Feldman (2002).

The complexity of a concept is given by the length of the Minimum Description Length
(MDL) code (Grünewald, 2007). The MDL code length of a rectangular concept, such as
the 2 (context) x 10 (shading) dimensioned concepts used in Chapter 7, is given by the
sum of the description length of a hypothesis, \( DL(H) \), and the description length of the
data described by that hypothesis, \( DL(D|H) \) (in other words, the description length of the
likelihood); hence, the full MDL code is given by \( DL(D|H) + DL(H) \). It is well known that
probabilities are equivalent to code lengths (see (Grünewald, 2007)); hence, we only need
to consider that \( DL(H) = -\log_2 P(H) \) and \( DL(D|H) = -\log_2 P(D|H) \) (the so-called
Shannon code) in order to compute the MDL complexity of the rectangular concepts.
Following Fass and Feldman (2002), we first compute the number of different rectangular
hypotheses that can 'fit' into the larger 2 x 10 concept space. There are 19 different possible
hypotheses (i.e., 1 x 1, 1 x 2,1 x 3,...,2 x 10). We assume that all possible rotations of a
hypothesis are sampled uniformly; hence, the full prior probability of a hypothesis, \( P(H) \),
is given by the likelihood of selecting a hypothesis of a certain shape, \( P_{m\times n} = 1/19 \), times
the number of possible ways that the hypothesis can fit into the category space, \( N_{m\times n} \).
For instance, a 1 x 2 rectangular hypothesis can be placed into the larger 2 x 10 concept
in 28 different ways; hence, the \( D(H_{1\times2}) = \frac{P_{1\times2}}{N_{1\times2}} = \frac{1}{19} \cdot \frac{1}{28} \). Each of the hypotheses are
drawn independently and can be combined to create non-rectangular shaped hypotheses;
the description length is then given by summing the $D(H)$ of the component hypotheses.

To compute the likelihood, $P(D|H)$ we consider the number of combinations of X data points that can be drawn from a particular hypothesis. Hence, the description length of the likelihood, is given by:

$$DL(D|H) = -\log_2 \left( \frac{1}{\binom{H\text{ size}}{n}} \right)$$

(12.1)

where $H_{\text{size}}$ refers to the area of the current hypothesis and $n$ is the number of data points and the denominator of the fraction in 12.1 is the number of permutations of $n$ data points drawn from area $H_{(\text{size})}$.

The complexity of the concepts used in Yang and Lewandowsky (2004) is computed in a similar fashion. We first assume that the category space can be discretized into a 900 (X pixels) by 900 (Y pixels) x 2 (context) space comprised of unit cubes\(^1\). The prior probability of any cuboid hypothesis is computed by first assuming that all hypotheses are equally likely and that for a given hypothesis, all possible unique rotations of that hypothesis within that category space are equally likely (see Figure 12.1 for examples of cuboid hypotheses).

The number of unique rotations of a cuboid hypothesis $(a \times b \times c)$ within in a larger cuboid space $(m \times n \times p)$ is given by:

\(^1\)These dimensions were derived from Figure 1 in Yang and Lewandowsky (2004). An identical category space was used in Lewandowsky et al. (2006)
For a 900 x 900 x 2 space there are 810,000 different cuboid hypotheses (hence, \( P_{m \times n \times p} = \frac{1}{810,000} \) ); hence the prior probability of a hypothesis equals \( \frac{P_{m \times n \times p}}{N_{m \times n \times p}} \) (cf., Fass & Feldman, 2002). Cuboid hypothesis are then populated throughout the space with the total \( D(H) \) equalling the sum of the component description lengths, and likelihoods are computed in the same fashion as 12.1 but substituting hypothesis volume for hypothesis area.
(a) Example category space.

(b) Example cuboid hypotheses.

Figure 12.1: Example cuboid category space and cuboid hypotheses for complexity analysis.