Journal of Experimental Psychology: General
2002, Vol. 131, No. 2, 163–193

0096-3445/02/$5.00 DOI: 10.1037//0096-3445.131.2.163

Simplified Learning in Complex Situations: Knowledge Partitioning in Function Learning

Stephan Lewandowsky, Michael Kalish, and S. K. Ngang
University of Western Australia

The authors explored the phenomenon that knowledge is not always integrated and consistent but may be partitioned into independent parcels that may contain mutually contradictory information. In 4 experiments, using a function learning paradigm, a binary context variable was paired with the continuous stimulus variable of a to-be-learned function. In the first 2 experiments, when context predicted the slope of a quadratic function, generalization was context specific. Because context did not predict function values, it is suggested that people use context to gate separate learning of simpler partial functions. The 3rd experiment showed that partitioning also occurs with a decreasing linear function, whereas the 4th study showed that partitioning is absent for a linearly increasing function. The results support the notion that people simplify complex learning tasks by acquiring independent parcels of knowledge.

The extensive literature on knowledge acquisition and representation cannot be summarized by a few simple assertions. Nonetheless, we suggest that it is possible to identify two principal themes that run through much of that literature: First, people's knowledge—particularly in areas of a person's expertise—is typically thought to be homogeneous and well-integrated (e.g., Bédard & Chi, 1992; Glaser, 1996). This theme is accompanied by the pervasive, if often tacit, corollary that prolonged learning and experience will resolve inconsistencies and contradictions in one's knowledge (e.g., Ericsson, 1996). The second principal theme puts boundaries on the first one by emphasizing the specificity of most knowledge. This theme is expressed by the ubiquitous twin findings that knowledge is typically limited to a specific domain (e.g., Ericsson & Charness, 1994) and that it can become inaccessible even after small changes to content or context (e.g., Reed, Dempster, & Ettinger, 1985; see Kimball & Holyoak, 2000, for a recent review).

This article critically examines the first theme: the presumed homogeneity of knowledge within a domain. We explore the implications of several recent reports that arguably pointed to a more heterogeneous and contradictory structure of knowledge, both in real-life tasks involving experts (e.g., Lewandowsky & Kirsner, 2000; Schliemann & Carraher, 1993) and laboratory tasks involving novices (e.g., Lewandowsky, Kalish, & Griffiths, 2000) and intermediates in an educational setting (e.g., Tirosh & Tsamir, 1996). In those studies, people were found to engage in contradictory behavior as a function of an objectively irrelevant context in which a domain-relevant problem was presented. Here, we focus on the nature of such seemingly heterogeneous knowledge and how contradictory behavior can emerge during early stages of learning under controlled conditions. In marked contrast to most previous studies on the contextualized nature of knowledge, our research involved extensive and interleaved training using all contexts that were presented at test.

We report four function–concept acquisition studies in which participants learned to predict a continuous response variable from a continuous stimulus variable. Training stimuli were accompanied by one of two context cues, and the principal manipulation was whether the context cue was systematically associated with components of the to-be-learned function. As discussed later, our main findings show that when the context cue was paired with different ranges of stimulus values, and hence predictive of the slope of a to-be-learned quadratic function, generalization was found to be context specific (Experiments 1 and 2). For example, when a context cue was primarily associated with a decreasing relationship between stimulus and response magnitudes during training, extrapolations outside the training range in that context tended to be influenced by the negative slope that predominated at training. This suggested that people use a context cue, when available, to gate the learning of simpler partial functions, a phenomenon we refer to as *knowledge partitioning*. Owing to the symmetry of the quadratic function, context did not directly predict absolute response magnitudes and hence was normatively irrelevant. In a third experiment, in which participants learned a linear function with a negative slope, the context cue predicted response magnitudes in addition to being systematically associated with stimulus values. People were again found to partition the common linear relationship into two parallel functions, each with a context-

Stephan Lewandowsky, Michael Kalish, and S. K. Ngang, School of Psychology, University of Western Australia, Crawley, Western Australia, Australia.

Preparation of this article was facilitated by Large Research Grant A79941108 and Discovery Grant DP0208035 from the Australian Research Council to Stephan Lewandowsky and Michael Kalish. We thank Tom Griffiths and Donald Rowe for their assistance during data analysis, and Ivonne Ambroz, Mike Mundy, and Leo Roberts for assistance during data collection.

Correspondence concerning this article should be addressed to Stephan Lewandowsky or Michael Kalish, School of Psychology, University of Western Australia, Crawley, Western Australia 6009, Australia. E-mail: lewan@psy.uwa.edu.au or kalish@psy.uwa.edu.au. Web site: http://www.psy.uwa.edu.au/user/lewan/

specific intercept. In a final experiment, a linear function was also used, but this time its slope was positive, which is known to coincide with people's a priori expectation of functional relationships and to be most readily learnable. Partitioning did not occur in this situation, and people disregarded context although it predicted response magnitude.

We interpret our results as showing that when people are confronted by a difficult learning situation, they seek ways to simplify the task by creating independent parcels of knowledge. Knowledge parcels are selectively accessed on the basis of any cue that can serve as a *gating* variable, even if that context cue is normatively unrelated to the response. Once accessed, knowledge parcels are used to determine a response without consideration of knowledge demonstrably present in other parcels. In contrast, when the learning situation is simple and conforms to prior expectation, people acquire a single knowledge parcel that captures the common underlying function and ignore context even when it normatively predicts response magnitude.

We conclude that knowledge can be partitioned and inconsistent within a well-trained problem domain. We suggest that context not only bounds the extent of one's knowledge but may also determine which of several mutually inconsistent and independent responses is elicited.

The structure of this article is as follows: We first seek support for our assertion that knowledge is most often viewed as being integrated within a domain but bounded in its extent. We then examine reports of inconsistent and contradictory behavior within a well-studied domain and present knowledge partitioning as an explanatory framework. We next identify the conditions necessary for a controlled examination of knowledge partitioning and justify the choice of function learning as a suitable experimental paradigm. The four experiments are then presented and their implications for current views of knowledge representation are examined.

Modal View of Knowledge: Bounded Integration

According to folklore, Isaac Newton formalized the concept of gravity while sitting under a tree when an apple fell on his head. Despite this very specific context of acquisition, there is no doubt that Newton applied the resulting concept of gravity with considerable generality, across many different contexts. Indeed, it seems preposterous to assume that Newton might have had different expectations about the behavior of, for example, oranges or olives falling off trees. Instead, we can safely assume that his knowledge of gravity was homogeneous and context-independent and thus applied to apples, oranges, and olives alike.

We suggest that this assumption of homogeneity and integration forms a major theme that links research into knowledge acquisition and structure across many domains. One may question the wisdom of any attempt to identify major themes within an area as extensive as research into knowledge acquisition—and indeed, to date, most research has proceeded independently and in parallel within several isolated domains. This is illustrated by the results of an on-line search using the PsycINFO database: Of the nearly 75,000 articles identified in response to the keywords *expert* and *knowledge*, only 3% contain both terms. Likewise, of the 33,000 articles on *expertise* and *problem solving*, only 2% share both keywords. Against this background, recent attempts to bridge some of those diverse domains by a discussion of common principles are particularly encouraging. For example, Kimball and Holyoak (2000) reviewed the nature of transfer in problem solving and expertise and successfully identified common attributes and principles. Fauconnier and Turner (1998) went further still by identifying several key notions that cut across many levels of analysis. For example, Fauconnier and Turner suggested that *frames* structure not only our conceptual and social life but also form the building blocks of basic knowledge by providing for the construction of word meaning. Fauconnier and Turner furthermore suggested that the same generality applies to constructs such as analogical mapping and reference points, among others. Finally, Frensch and Buchner (1999) identified common strands in the debates on whether human cognition is domain-specific or domain-general across a range of areas, including, in particular, expertise and human development.

Encouraged and emboldened by these precedents, we chose to place our experiments in a wide and general context. Accordingly, this section first identifies the integrated nature of expert knowledge before turning to representative research in the domains of category learning and skill acquisition. Next, we consider findings across several domains that have consistently put boundaries on the extent to which knowledge is integrated. The consensus of those findings is that knowledge can become inaccessible if tested outside the context or content of its acquisition. Finally, we focus on recent reports of apparent exceptions to the integration assumption *within* the context in which knowledge has been acquired; it is those exceptions that our experiments further explore.

Integration and Homogeneity

We adopt the working definition of knowledge offered by Charness and Schultetus (1999) as "acquired information that can be activated in a timely fashion in order to generate an appropriate response" (p. 61). This definition entails two implications: First, we are not concerned with the performance of people whose knowledge is demonstrably imperfect or sketchy or based on an uncertain acquisition history prior to commencement of formal training (such as intuitive physics; e.g., Cooke & Breedin, 1994). Second, we must necessarily be concerned with the nature of expertise, because expert performance is widely thought to rely primarily on domain knowledge rather than exceptional computational ability (e.g., Charness & Schultetus, 1999). An example of this is that the knowledge base of world chess champion Garry Kasparov has been estimated to contain some 100,000 memorized chess configurations (Charness & Schultetus, 1999).[1]

Expert Knowledge

Research on human expertise is permeated by the assumption that expert knowledge is characterized by high levels of organization and integration. This view has been stated in a variety of forms. For example, Bédard and Chi (1992) claimed that "experts have more and stronger links among concepts [than novices], suggesting that there is a greater degree of connectedness and

[1] A distinction has been made between two types of expertise: routine and adaptive (Kimball & Holyoak, 2000). The former refers to performance based primarily on familiar domain-specific tasks (e.g., playing chess) and is of primary concern here. The latter refers to domains or problem-solving styles by individuals that are inherently more variable and involve less stereotyped solutions. We do not consider adaptive expertise in this article.

cross-referencing, and the pattern of connections and cross-referencing can result in a better structure" (p. 136). A stronger version of the assumption was formulated by Glaser (1996), who pointed to 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. (pp. 305–306)

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).

The hallmarks of integrated knowledge, then, are the existence of extensive links between sometimes disparate concepts, a large number of records of previous experiences, and the ability to coherently apply that knowledge to many different problems. Integration does not preclude the possibility that knowledge may be applied by different routes, using one of several complementary coexisting strategies. Indeed, there is much evidence for the coexistence of alternative strategies at all levels of skill acquisition and expertise (e.g., Lovett & Schunn, 1999; Shrager & Siegler, 1998); we examine this issue in the General Discussion. However, integration does entail the expectation that those alternative strategies cooperate rather than compete, thus ensuring that responses are coherent rather than contradictory, even if driven by different strategies.

If expert knowledge within a domain is integrated in this manner, one visible consequence should be the occurrence of spontaneous transfer between similar problems: That is, if an expert demonstrates the ability to apply a known solution to a problem, then a similar problem can be expected to elicit the same route to solution, provided the person's knowledge is sufficiently integrated to cross-reference the various trigger conditions for a known solution (cf. Chi, Feltovich, & Glaser, 1981).

There is considerable support for the occurrence of within-domain transfer: Novick and colleagues (Novick, 1988; Novick & Holyoak, 1991) have shown that mathematical expertise predicts the extent to which people spontaneously transfer solution strategies from one algebra word problem to another that shares the same deep structure, but not surface features, even when the two are administered under two separate experimental cover stories. In one study, the extent of transfer among expert participants was found to be up to nine times greater than that among novices (Novick, 1988, Experiment 1), in line with the view that expert knowledge can be considered highly integrated.

In the domain of accounting, Marchant, Robinson, Anderson, and Schadewald (1991) similarly found that experts, by default, show significantly more transfer between problems involving the application of taxation laws than do novices. This finding was accompanied by the additional fact that when processing of the source problem was enhanced (e.g., by predicting a legal opinion before it was revealed or by studying two related problems), the experts' subsequent transfer was reduced to the level shown by novices. Marchant et al. argued that this seemingly paradoxical reduction in transfer arose because all problems constituted exceptions to a general taxation principle. Hence, enhanced 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). For present purposes, this interpretation underscores the integration of expert knowledge, as it shows that even exceptional cases can trigger the knowledge embodied in general rules. (The fact that this may not be in the expert's best interest is beside the point here.)

Knowledge integration is also a quintessential aspect of expertise in medical diagnosis. Unlike many other domains, medical expertise is based on two distinct categories of knowledge: namely, clinical expertise incorporating diseases and symptoms and knowledge of basic biomedical science (Patel, Arocha, & Kaufman, 1999). Conventionally, curricula first focus on the acquisition of basic biomedical knowledge, which is followed by the acquisition of clinical knowledge. Because both types of knowledge are essential to successful medical diagnosis, with clinical reasoning being expected to rest on a foundation of scientific knowledge, their integration is particularly important (Patel et al., 1999).

Much of the empirical evidence for the integration of biomedical and clinical knowledge derives from the fact that increasing medical expertise is nonmonotonically related to the overt application of biomedical knowledge. That is, in a simulated diagnosis situation, the think-aloud protocols of senior medical students contain more propositions reflecting biomedical knowledge than the protocols of beginning medical students, but—seemingly paradoxically—the protocols of experienced physicians, like those of the beginners, contain little mention of biomedical facts (e.g., Boshuizen & Schmidt, 1992). Because the advanced experts could be prompted at a later point to provide an extensive biomedical analysis of their diagnosis, whose level of detail surpassed that offered by the medical students, the experts' decreased mention of biomedical knowledge clearly did not reflect its loss or inaccessibility. The further fact that the correlation between diagnostic and subsequent analytic protocols, as measured by the number of shared propositional arguments (i.e., biomedical keywords), increased monotonically with expertise confirmed that biomedical knowledge was essential to diagnosis at all levels of expertise, although its role became increasingly more tacit (Boshuizen & Schmidt, 1992).

It follows that the decreasing overt mention of biomedical knowledge that accompanies increasing expertise actually reflects its growing integration with clinical knowledge. This process, labeled knowledge encapsulation, has been described as the "progressive subsumption, or packaging, of lower level concepts . . . under a limited number of high-level concepts with the same explanatory power" (Schmidt & Boshuizen, 1993, p. 347). Ultimately, as a result of this process, "expert medical knowledge is organised in a large and highly coherent knowledge network" (van de Wiel, Boshuizen, & Schmidt, 2000, p. 349). The integration is particularly noteworthy because it spans two distinct categories of knowledge within an expert domain.

Finally, the primacy of integration of knowledge is reflected in current theories of expert performance. Specifically, the EPAM-TEMP theory of Gobet and Simon (e.g., 1996b) suggests that experts, for example, master chess players, organize their knowledge in terms of templates. Templates

> are patterns of the chess board found in familiar openings and lines of play, again of the sorts frequently encountered in games. The templates specify the locations of perhaps a dozen pieces in the position

> (thus specifying a class of positions), but also contain variables (slots) in which additional information can be placed. (p. 29)

Most relevant here is that this additional information includes links to other related templates, such as board configurations anticipated later in the game, thus creating a richly interconnected net of domain knowledge.

Skill Acquisition and Categorization

The integration theme also permeates other areas of inquiry, although perhaps with less force than in expertise research. Thus, some theories of skill acquisition assume that different forms of knowledge can support task performance: Logan's (e.g., 1988) instance theory postulates that skill acquisition consists of the transition from the execution of an initial slow algorithm to the fast retrieval of memorized earlier solutions, and similar transitional views have been advanced by several other theorists (e.g., knowledge compilation, J. R. Anderson & Fincham, 1994; component power law, Rickard, 1997). Common to most of these views is the unidirectionality and finality of the transition between different forms of knowledge: It is assumed that people's initial approach to the task is gradually abandoned in favor of a maturing alternative that eventually provides the ongoing and sole basis of performance. Once the transition to skilled performance has been completed, the underlying knowledge is again often considered integrated. For example, Müller (1999) postulated that skilled performance, such as comprehending a computer language, depends on the "integration of features to form a conceptual knowledge unit" (p. 194). Although he left the nature of such knowledge units largely unspecified, Müller provided some experimental support for the integrative nature of knowledge in preference to a more task-specific view.

In categorization, conventional theorizing has at least tacitly relied on integration of knowledge. This is readily understood for rule-based (e.g., Ashby & Gott, 1988) or prototype-based (e.g., Homa, Sterling, & Trepel, 1981) approaches, which by definition condense disparate experiences into a single integrated piece of knowledge—either a rule or a prototype—that determines categorization performance. The integration theme may appear less obvious with instance models of categorization, such as Nosofsky's (e.g., 1991) GCM, which retains identifiable representations of all stimuli presented during training. These models nonetheless arguably instantiate integration because all representations are considered during a categorization decision, and because all representations are treated qualitatively identically. That is, although different stimulus dimensions may receive differing attentional weights (e.g., Kruschke, 1992), a test stimulus is compared with all previously encountered instances from all learned categories.

Status of the Integration Theme

We conclude that, on balance, the evidence and available theorizing favor a view that leans toward integration and homogeneity of knowledge. This is most strongly evident in the area of human expertise, but the theme has also been expressed in much theorizing in categorization and, to a lesser extent, skill acquisition. Although this conclusion rests on a review that necessarily had to remain selective, the large number of supportive reports contrasts sharply with the scarcity of exceptions, which we discuss later.

Limits to Integration: Content and Context

Virtually all claims of integration are accompanied by the realization that the scope of integration is limited. There is near unanimity that knowledge is specific to a domain or situation and that it may become inaccessible or inappropriate if the problem context is altered. At this juncture, the relevant findings are sufficiently general to permit a loose usage of *context* as referring to the characteristics of a problem and its materials or the environment in which it is tested; later, when we present our experiments, *context* will refer more narrowly to a cue that accompanies a problem.

In research on expertise, there is no doubt that whenever context is changed such that a problem is taken outside the expert's domain, however narrowly defined, performance greatly deteriorates. Classic examples of this deterioration arise when a domain-relevant problem is made atypical, for example, by presenting chess masters with a board for memorization that contains randomly arranged chess pieces, as opposed to midgame or end-of-the-game configurations, for which expert memory is far superior (e.g., Gobet & Simon, 1996b). Similar effects have been observed in many other domains: A review by Ericsson and Lehmann (1996) cites areas as diverse as the games of bridge, go, othello, snooker, basketball, field hockey, volleyball, and football, and professional disciplines such as medicine, computer programming, and dance.[2]

Outside the area of expertise, research on general problem solving is similarly characterized by the frequent lack of transfer between problem isomorphs. Changes in cover story (e.g., military vs. medical; Gick & Holyoak, 1980) or instrument (e.g., laser vs. ultrasound; Holyoak & Koh, 1987) are sufficient to prevent people from spontaneously transferring a solution strategy from one problem to another. In addition, what little transfer may occur is further reduced or eliminated by a shift of environmental context (e.g., classroom vs. laboratory; Spencer & Weisberg, 1986). In research on skill acquisition, the specificity of skills and the elimination of transfer even after subtle changes to the stimulus material (e.g., Logan & Klapp, 1991) are well-known. In categorization research, people have been shown to be able to transfer their knowledge of a category to a novel object when given the appropriate category label; however, once a novel object has been thus labeled, people are reluctant to transfer further knowledge from additional categories unless specific mappings are provided (Moreau, Markman, & Lehmann, 2001).

---

[2] In several of these cases, follow-up research has demonstrated that experts' memory for atypical stimuli is still better than that of novices (e.g., Gobet & Simon, 1996a). There are also domains in which memory performance does not monotonically increase with expertise (e.g., medical diagnosis; Schmidt & Boshuizen, 1993). None of these exceptions detracts from the main point that expert memory is highly sensitive to transfer to atypical stimuli. The study of expert memory has also been criticized for involving tests that are said to be contrived, because they may not tap the memory that is essential to task performance (Vicente & Wang, 1998). These criticisms were met rather convincingly by Ericsson, Patel, and Kintsch (2000).

Overall, although shifts in context or content can give rise to an intricate pattern of results (Kimball & Holyoak, 2000), most of the research can be characterized by two principal features: First, the predominant methodology involves transfer from one context at study to another, novel context at test. Second, the research has primarily delineated the extent of integrated knowledge and has underscored its domain specificity, without questioning its homogeneity within a domain.

The next section analyzes a representative sample of exceptions to this apparent consensus. We focus on reports of violations of the homogeneity assumption within domains that one would expect to be characterized by integration.

Violations of Homogeneity: The Case for Knowledge Partitioning

Heterogeneity in Categorization

Natural categories, embodying knowledge of concepts such as dogs, animals, or furniture, have traditionally been thought of as relatively stable, static, and context-independent (Medin & Ross, 1989). This conventional view has been challenged by several observations that the use of natural categories is context specific. For example, consider which beverage would be more typical of a librarian taking a morning break: Tea or milk? Now suppose it was a truck driver taking a break: Would tea still be more typical? Roth and Shoben (1983) showed that people will choose one or the other beverage as being more typical, depending on the occupational context in which the question is presented. Roth and Shoben suggested that "these results argue strongly against the existence of an invariant semantic space" (p. 370).

For classification learning tasks, two recent theories of categorization have incorporated two coexisting but separable types of knowledge: an overarching rule that permits classification of most, but not all, stimuli, and memorization of specific instances, or parts of instances, that violate that rule (e.g., Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994). Numerous experiments have adduced support for the rule-plus-exception view, for example, by showing that generalization to novel stimuli is governed by one or the other process, depending on the relative proximity of a novel item to exceptional stimuli (Erickson & Kruschke, 1998; however, see Nosofsky & Johansen, 2000).

Experimentation has also identified conditions under which people may acquire seemingly independent types of knowledge, for the same set of stimuli and during a common learning phase, that can then be selectively brought to bear in response to differing task demands. Whittlesea, Brooks, and Westcott (1994) directed people to treat training stimuli *analytically* (by emphasizing the relevance of features to categorization) and, in a separate learning phase, *holistically* (by promoting integral encoding of each training item). Across five experiments, people demonstrably used one or the other type of knowledge, as identified by generalization patterns, depending on task demands at test. The distinction between holistic and analytic knowledge of categories was independently supported by Goldstone (1996), although his comparisons were across groups of participants. (Goldstone, 1996, preferred the terms *isolated* and *interrelated* to *holistic* and *analytic*, respectively, but both sets of terms describe the same dichotomy.) Together, these studies show that categorization can involve seemingly distinct types of knowledge that are differentially created or triggered, depending on study conditions and test demands.

In a similar vein, Aha and Goldstone (1992) showed that people could learn a categorization problem by applying different weightings to predictors in different regions of the stimulus space. Aha and Goldstone sampled training stimuli from two distinct clusters in a two-dimensional category space, each bisected by its own uniquely oriented boundary. The data suggested that people create a highly heterogeneous representation of the stimulus space, relying on a separate decision boundary for each region, with generalizations to new stimuli based on the proximity of a transfer item to the closest available boundary.

In summary, both naturalistic and laboratory-trained forms of categorical or conceptual knowledge can be heterogeneous and potentially contradictory. Context may determine one's perception of typicality or category membership, and stimulus neighborhood may influence one's classification response. As it turns out, even more extreme forms of heterogeneity can be observed in expert knowledge.

Expertise Is Not Always Integrated

Perhaps because the dominant theoretical tenet expects expertise to be characterized by notable integration, cross-referencing, and consistency, there has been little theoretical impetus to search for inconsistencies or contradictory knowledge among experts. Accordingly, only a few reports of context-bound contradictory behavior are known to us, but they represent perhaps the most striking instances of heterogeneity available.

We first look at an examination of knowledge that, although 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 their participants correctly identified two infinite sets as containing the same number of elements, the vast majority of the same students (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 that 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 participants were highly conversant with both representations of all functions used in the study, the findings by Even point toward heterogeneity even in consolidated knowledge.

Contradictory elements of expertise 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 everyday 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 in either a pricing context ("If 2 kg of rice cost 5 cruzeiros, how much do you have to pay for 3 kg?"; Schliemann & Carraher, 1993, p. 60) or 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?"). It is important to note that both problem contexts were familiar to participants and relevant to their domain of expertise. Schliemann and Carraher (1993) 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, participants 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, in contrast, accuracy was dramatically lower (20%), and half of the responses given were based on estimation.

The foregoing studies share two potential limitations: First, in all cases, it is debatable whether participants could be truly considered experts by commonly used performance criteria (e.g., Ericsson & Charness, 1994). Second, the 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. These limitations do not apply to the next study.

Lewandowsky and Kirsner (2000) asked experienced wildfire commanders to predict the spread of simulated wildfires. The spread of wildfires is primarily determined by two physical variables: All other factors being equal, 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 under those circumstances, the experts' predictions depended 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 both apples and oranges obey the laws of gravity.

Before presenting an explanatory framework for these results, it is essential to differentiate them from the conventional context effects reviewed earlier. Four attributes of the Lewandowsky and Kirsner (2000) study are relevant in this regard: (a) 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. (b) 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. From now on, we restrict consideration to this very local and subtle meaning of *context* as a label that accompanies a problem. (c) Both domain-relevant contexts were part of the training regime of the experts, and both regularly occurred in the field. That is, unlike the conventional transfer methodology, in which knowledge acquired in one context is tested in another novel one, both test contexts were interleaved throughout training. (d) The context shift resulted not only in a reduction of performance, as for example observed with chess masters' memory of random board configurations, but also 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. (Because stimuli were arbitrarily created, no assessment of accuracy was possible.)

We suggest that these attributes are sufficiently unique to warrant exploration of an explanatory framework for the results reviewed in this section. This framework, *knowledge partitioning*, asserts that knowledge, even within a well-learned domain, can be anything but homogeneous.

Knowledge Partitioning: A Framework for Heterogeneity

Lewandowsky and Kirsner (2000) explained their finding, that highly skilled experts would make two mutually opposing predictions under identical circumstances and identical surface structure, by proposing the construct of knowledge partitioning. According to knowledge partitioning, expertise—and by plausible implication, other less consolidated forms of knowledge, too—is best viewed as being partitioned or modularized, with the strong possibility that inconsistent or mutually exclusive components of knowledge can persistently coexist in independent parcels. Thus, task-relevant knowledge is tied to the context of its acquisition and use, without necessarily being cross-referenced to other preexisting or coacquired knowledge structures. In consequence, expert predictions of back burns may run completely counter to predictions for equivalent to-be-controlled fires. On an extreme version of that view, Newton might have predicted that oranges soar from trees whereas apples fall down and onto people's heads.

The knowledge partitioning framework can account for the results of Lewandowsky and Kirsner (2000) and other instances of contradictory expert behavior (e.g., Schliemann & Carraher, 1993). However, as currently stated, the view entails two major problems relating to identifiability and demarcation of knowledge parcels, respectively.

Identifiability of Knowledge Parcels

The first problem is that in the absence of an independent means of identifying knowledge parcels, the view is potentially circular. That is, an observed contradictory behavior is explained by invoking constructs (i.e., knowledge parcels) that are identified by the same behavior. It is difficult to sidestep this problem when examining the performance of experts, as their lifelong history of learning is neither knowable nor controllable. Instead, the identification problem is best solved by observing the acquisition of knowledge in the laboratory, where partitioning can be experimentally encouraged and controlled.

Lewandowsky et al. (2000) pursued this option in a category learning experiment that involved a laboratory analog of wildfire prediction. There is a long-standing link between research in expertise and categorization (e.g., Medin & Edelson, 1988); accordingly, Lewandowsky et al. (2000) reduced the wildfire task to a categorization problem in which two diagnostic variables (wind strength and slope gradient) and a situational context (to-be-

controlled fire or back burn) predicted a category (uphill vs. downhill fire spread). Training stimuli were arbitrarily constructed such that context provided a simple but imperfect classification rule: Nearly all back burns spread uphill into wind, and most to-be-controlled fires spread downwind and downhill. Lewandowsky et al. found that in most circumstances, participants relied primarily on the simple context rule, together with memorization of a few exceptions, at the expense of the correct (but more complex) true category boundary formed by a nonlinear combination of wind and slope. Thus, participants came to resemble the experts, by predicting that back burns would spread uphill and to-be-controlled fires downwind, giving only partial consideration to the effects of wind strength and slope gradient.

One interpretation of these results is consistent with the knowledge partitioning framework offered for the expert data. In that view, participants formed two independent parcels of knowledge about the task: one that captured knowledge of back burns and another one that accommodated to-be-controlled fires. Each parcel can be thought of as containing a strong link to one outcome—uphill or downhill spread, respectively—plus an attenuated representation of the effects of wind and slope. Parcels of that type would be compatible with the differing representations of different regions of the stimulus space in the study by Aha and Goldstone (1992) mentioned earlier. This interpretation would also underscore the generality of the knowledge partitioning framework, and it would suggest that parcels can emerge after relatively little training. Moreover, unlike the original account of the expert data offered by Lewandowsky and Kirsner (2000), the results of Lewandowsky et al. (2000) sidestep the circularity problem for two reasons. First, the anticipated parcels were predefined by the design of the stimulus space, and second, given the nature of the test items, the rules used by each individual participant were identifiable by statistical means. The experiments below adopted a variant of this approach and thus also avoided the identifiability problem. However, the experiments by Lewandowsky et al. (2000) did not resolve the second problem that has thus far been associated with knowledge partitioning: namely, that in some cases, the observed contradictory behaviors may be more apparent than real.

Apparent Contradictions May Not Be Real

Consider again the studies by Schliemann and Carraher (1993) and Lewandowsky and Kirsner (2000). The interesting finding in those studies was that problem context affected performance despite being physically irrelevant and despite there being a single context-invariant correct solution. That context invariance, however, may have been obscured by coincidental features of the task environment. For example, back burns differ from to-be-controlled fires in several key aspects: Being lit and managed by fire fighters, back burns rarely roar out of control, they tend not to be used in strong wind conditions, and they tend to be lit in the proximity of natural fire breaks. This renders the experts' experience with back burns noticeably and consistently different from that with to-be-controlled fires. Moreover, to-be-controlled fires, which are invariably more powerful than back burns, can create their own convective winds that may render their local behavior different from smaller fires under identical topographical and meteorological conditions. These verifiable differences between problem contexts may engender the learning of strong associations between the type of fire and its expected behavior, which in turn may completely obscure the operation of uniform, context-invariant physical variables.

In a similar manner, in the study by Schliemann and Carraher (1993), people may have adjusted the grain size with which they report their answer while relying on the same integrated underlying knowledge. Goldsmith, Koriat, and Weinberg-Eliezer (2002) have shown that "when people report information from memory, they use control over grain size in a calculated manner, attempting to strike a balance between the goal of being accurate and the goal of being informative" (p. 88). That is, depending on the balance of incentives, people might choose to recall the time at which they experienced an event as 5:23 p.m. (informative, but easily inaccurate) or between 5 and 5:30 (likely to be correct, but less informative). The cooks studied by Schliemann and Carraher (1993) may have responded in an analogous manner to subtle incentives that differed between the recipe and pricing contexts.

Related concerns apply to the categorization study by Lewandowsky et al. (2000), in which context itself was a valid (albeit imperfect) predictor. Because the issue is fundamental to the present experiments, we create a distinctive nomenclature. We refer to situations in which a context cue, by itself, is statistically predictive of an outcome as a *first-order* effect of context. Critically, whenever context has a first-order relationship to an outcome, the occurrence of contradictions need not reflect the creation of separate parcels of knowledge. In those situations, contradictions may instead arise when context is incorporated in a single multidimensional decision boundary that also considers other predictor variables. For example, in the study by Lewandowsky et al., people may have learned to orient a single decision boundary through a three-dimensional categorization space defined by the three predictors of context, wind, and slope. This may give rise to apparent contradictions, because stimuli with identical wind and slope values may lie on opposite sides of a decision boundary when they are separated along the third, context dimension.

However, use of an integrated multidimensional decision boundary clearly would not constitute knowledge partitioning. Thus, the results of Lewandowsky et al. (2000) may be another case in which the observed contradictions, although compatible with the idea of knowledge partitioning, were more apparent than real. A similar comment applies to the heterogeneity observed with natural categories (e.g., Roth & Shoben, 1983): It is possible that the context in which one observes beverages being consumed throughout one's life forms an integral part of one's knowledge about beverages. Hence, if one's experience includes the fact that tea and milk are primarily consumed by librarians and truck drivers, respectively, the type ordering between tea and milk might reverse across different contexts, without this reflecting a true contradiction in knowledge.

It turns out that contradictions can be identified as real, beyond being merely apparent, only if context is not directly predictive of an outcome but instead identifies the relationship between other predictor variables—a relationship that we term *second order*. A second-order relationship between context and a to-be-learned outcome gives rise to a chain of implications: First, because context by itself is not predictive, it cannot be incorporated into a single multidimensional decision boundary. Second, the absence of a first-order relationship renders it possible (and indeed likely) that people ignore context entirely, in the same way that people

typically place little value on nondiagnostic features during categorization and problem solving (Goldstone, 1996; Kruschke & Johanson, 1999; Lovett & Schunn, 1999). Third, and most crucial, if people nonetheless turn out to be sensitive to context, they must necessarily use it to gate access to separate components of knowledge. This would constitute strong evidence for the concept of knowledge partitioning, and our experiments were designed to discover such second-order effects of context. We further clarify the notion of gated access after the first experiment has been presented.

Examining Knowledge Partitioning: Choice of Paradigm

The goal of the present experiments was to examine the possible emergence of contradictory behaviors in naive participants under controlled laboratory conditions, characterized by the following features: During training, (a) all problems were characterized by a common context-invariant solution, (b) problems were presented in all contexts used at test in an interleaved manner, and (c) context bore a second-order relationship to the to-be-learned outcome. At transfer, (d) all problems were presented in both contexts; thus, knowledge partitioning would be said to be present if people's transfer responses to a given problem differed between contexts. In addition, the first two experiments retained a limited degree of surface similarity to the expert domain in which knowledge partitioning was first observed: namely, the behavior of wildfires. However, none of the participants had any formal task-relevant background knowledge. A function learning paradigm was considered to be most suitable for implementation of these methodological features.

Function Learning

In function learning, people learn the relationship between a continuous stimulus dimension and a continuous response variable from a set of discrete training items. On each learning trial, a particular stimulus value is presented, and the participant's task is to predict the associated response magnitude. Each response is followed by corrective feedback. All magnitudes are typically coded graphically and without any explicit numeric representation. For example, the stimulus value might be represented by an horizontal arrow of varying lengths, and participants might have to adjust a vertical slide rule with the mouse to indicate their response. Corrective feedback would be presented on the same slide rule. Owing to the absence of numeric coding, participants must learn the function in a purely abstract, cognitive manner.

With sufficient practice, people are remarkably adept at generalizing the discrete trials encountered during training into a continuous function embodying the underlying relationship (e.g., Busemeyer, Byun, DeLosh, & McDaniel, 1997). This is evident during a transfer phase when novel stimulus values are presented and participants must produce the associated response magnitudes. If those novel stimuli fall within the range of training values, thus requiring interpolation, performance is typically highly accurate (see Busemeyer et al., 1997, for a review). If novel stimuli are presented that are outside the range of training values, participants are also capable of extrapolation, albeit generally at a lower level of accuracy than interpolation (DeLosh, Busemeyer, & McDaniel, 1997). The ability to extrapolate varies considerably between different functions, with extrapolation being nearly perfect for linear functions but less accurate for exponential and quadratic functions and being generally better for positive (ascending) functions than for negative (descending) ones (DeLosh et al., 1997). Several theories have been put forward to explain these benchmark results, and at present, it appears that an approach based on generalization from memorized instances provides the best available explanation (DeLosh et al., 1997). We resume discussion of relevant theories after the experiments have been presented.

Function Learning and Knowledge Partitioning

Our focus on possible knowledge partitioning implies that a conservative choice of paradigm would be one that is known to lend itself to integration. There is no question that function learning provides an integrative methodology par excellence. One indication of this is people's ability to abstract a coherent function from numerous discrete stimuli that can be used for interpolation and extrapolation. The fact that extrapolation can be nonmonotonic (e.g., people can project a sine wave beyond the trained range) is particularly diagnostic of the abstraction and integration across instances that must have occurred during training (Bott & Heit, 2001). The integrative nature of function learning is also apparent in Müller's (1999) choice of the paradigm as being closely related to his hypothesis of conceptual integration in skill acquisition. It follows that, should partitioning occur in a function learning paradigm, it would emerge against a background of pervasive integration.

In addition, function learning also lends itself to implementation of a second-order—without first-order—relationship between context and the to-be-learned response. Specifically, in Experiments 1 and 2, participants learned concave quadratic functions. To facilitate the partitioning of knowledge during training, we ensured that the two levels of a binary context variable were probabilistically associated with different ranges of stimulus values. Thus, one context was primarily associated with stimuli taken from the increasing component of the quadratic function, and the other was primarily associated with stimuli taken from the decreasing component. Because of the symmetry of the quadratic function, this ensured that context did not predict response magnitude and hence bore no first-order relationship to the to-be-learned responses. Nonetheless, context stochastically identified the local slope of the quadratic function, thus implementing a second-order relationship. At transfer, stimuli from the entire range of stimulus values were presented in both contexts without corrective feedback. If people partitioned their knowledge during training, their performance at transfer should be affected by context. Responses for a given stimulus should differ between contexts although only one response magnitude was reinforced during training and although context did not directly predict a response.

Experiments 3 and 4 used a similar design, with a linear function whose slope was either negative (Experiment 3) or positive (Experiment 4). In contrast to the first two experiments, context did not identify slope but, because its two levels were preferentially paired with high and low stimulus values, context partially predicted response magnitude and thus had a first-order relationship with the to-be-learned response. These experiments tested the boundary conditions of the partitioning that turned out to emerge in the first two studies.

In all experiments, context was represented by a verbal label that accompanied each trial and by the color of certain components of the stimulus display. No other surface features of the stimuli differed between contexts. The subtlety of this manipulation may give rise to the concern that it might be ineffective or insufficient. However, there are numerous precedents in which subtle context cues have had striking effects. For example, in recognition memory, performance is greatly affected by whether the test item *jam*, for example, is accompanied by the context cue *traffic* or *strawberry* (Light & Carter-Sobell, 1970). The same has been shown to be true for implicit tests of memory (e.g., Lewandowsky, Kirsner, & Bainbridge, 1989, Experiments 4 and 5). Indeed, Tulving's (e.g., Tulving & Thomson, 1973) influential encoding specificity principle rests entirely on the mnemonic function of single study and retrieval cues. Finally, in situations related to the present paradigm, single-word context cues have been demonstrably effective (e.g., Lewandowsky et al., 2000; Lewandowsky & Kirsner, 2000). It follows that there is no reason similar cues cannot be effective here—indeed, the success of a manipulation as subtle as ours would underscore the ease with which partitioning of knowledge can be induced during training.

Experiment 1

The purpose of Experiment 1 was to examine whether knowledge partitioning may occur in nonexpert participants learning a moderately complex task. Because the idea of knowledge partitioning was originally stimulated by research on wildfire experts, the experiment used an abstract analog of the fire prediction task. Participants had to learn a function that related speed of fire spread (irrespective of direction of spread) to wind speed. Wind direction was assumed to be downhill throughout and slope was assumed to be constant. The chosen function was an upward concave quadratic that, although arbitrarily parameterized, resembled the true physical situation. The vertex of the function represented the wind speed at which the direction of the fire reversed: Above that point, the force of wind was sufficient to overcome the effect of slope, thus blowing the fire downhill with increasing speed. Below that point, the wind was insufficient to blow the fire downhill, and speed of spread uphill increased with decreasing wind strengths.

In addition to information about wind strength, training stimuli in Experiment 1 also contained information about the context of each fire; that is, whether it was a back burn or a to-be-controlled fire. In the systematic-context condition, most back burns were associated with slow wind speeds (and hence the decreasing part of the quadratic function), and most to-be-controlled fires were associated with high wind speeds (and hence the increasing component of the quadratic). In the randomized-context condition, in contrast, training stimuli across the entire range of the function were presented in both contexts. For comparison purposes, two additional conditions were included in which only the decreasing (left-only condition) and the increasing (right-only condition) components of the function were presented during training. The nature of people's knowledge was assessed during a transfer test that was common to all conditions and involved presentation of the entire range of stimulus magnitudes in both contexts.

On the basis of standard function learning results (e.g., DeLosh et al., 1997), participants in the randomized-context condition were expected to learn the complete quadratic function and to be able to do so with considerable accuracy. On the basis of knowledge partitioning, participants in the systematic-context condition were expected to learn the two halves of the function separately, so that during a subsequent transfer test, their extrapolation would be heavily influenced by a negative relationship between wind speed and fire spread in the back-burning context and by a positive relationship in the fire-fighting context.

Method

Design

The experiment involved a 4 (training condition) × 2 (test context) × 37 (stimulus value) between–within-subjects design. The only between-subjects variable was training condition, which manipulated the relationship during training between context and wind speed. In the systematic-context condition, context was predictive of wind speed such that low wind speeds primarily occurred in the back-burning context, whereas high wind speeds primarily occurred in the fire-fighting context. In the randomized-context condition, in contrast, all wind speeds were equally likely to occur in either context. In the left-only condition, most trials involved low wind speeds, and training thus focused on the descending part of the quadratic function. All training stimuli in that condition were shown in the back-burning context. In the right-only condition, most stimuli were high wind speeds taken from the ascending part of the function, and all stimuli were shown in the fire-fighting context.

Regardless of training condition, all participants were given a common transfer test in which all stimuli were presented twice in both contexts.

Participants

Fifty-two undergraduates from the University of Western Australia participated voluntarily in exchange for course credit. An equal number of participants were randomly assigned to each of the four conditions. Data from one participant in the right-only condition were lost because of equipment failure.

Apparatus and Stimuli

The experiment was controlled by an IBM-compatible computer that presented stimuli and collected responses. The to-be-learned function was an arbitrarily parameterized quadratic equation relating fire speed (F) to the two predictors, wind (W) and slope (S): $F = 0.2 \times [0.5 \times (W - S)]^2 + 8$. Wind direction always opposed slope, and S was arbitrarily set to 18 throughout, so that $F = 24.2 - 1.8W + 0.05W^2$. The vertex of the function represented the point at which the force of the wind balanced the effect of slope: in this case, when $W = 18$. To the left of that point, fires were slope driven, and their speed thus decreased with increasing wind speed. To the right, fires were wind driven, and their speed thus increased with increasing wind speed. Thirty-six training stimuli were used, ranging from wind speeds of 0 to 36, with 18 (the vertex of the function) omitted. At test, a wind speed of 18 was included, resulting in a total of 37 transfer stimuli.

Each stimulus graphically represented wind and context. Wind was represented by a horizontal arrow at the top of the screen, with the length of the arrow indicating the particular wind speed. No numerical value for wind speed was provided. The color of the arrow (blue or red) coded the context in which the fire should be considered. The assignment of color to context alternated between participants. Context was additionally identified by a textual label (i.e., *Fire-Fighting* or *Back-Burning*).

Participants were asked to predict the speed of the fire regardless of its direction of spread by using the mouse to drag a pointer from the far corner of the display and drop it on a vertical scale. The scale was labeled *slow spread* at the bottom and *fast spread* at the top, without any incremental values or tick marks. The use of a vertical response scale with a horizontal

stimulus and the absence of incremental values were aimed at preventing direct visual mapping from stimulus to response. A snapshot of the screen display during a learning trial is shown in Figure 1.

During training, a response was immediately followed by feedback, which consisted of an arrow positioned next to the vertical scale to indicate the correct magnitude. Predictions deviating by 5 or more units (approximately 2 cm on a 17-in. [43.2-cm] display monitor) from the correct answer were signaled by a tone indicating poor performance. Participants were required to acknowledge feedback by a mouse click. During the transfer test, feedback was absent. Stimuli were separated by a 2-s blank period, and the textual-context label preceded the remainder of each stimulus by 1 s.

Procedure

During training, each stimulus was presented five times. Thus, the systematic-context and randomized-context conditions included 180 training trials (five replications of each of 36 stimuli). In the left-only and the right-only conditions, there were 90 training trials, because training was almost exclusively restricted to one half of the function. The order of presentation of the stimuli was randomized separately for each participant, subject to the constraint that all stimulus magnitudes had to occur once within each block of 36 trials (18 trials for the left-only and the right-only conditions).

In the systematic-context condition, 90% (162 out of 180) of all slope-driven fires (i.e., $W < 18$) occurred in a back-burning context, and 90% of wind-driven fires (i.e., $W > 18$) occurred in the fire-fighting context. The remaining 10% (18) of trials were exceptions to this rule (i.e., fire-fighting context, with $W < 18$, and back-burning context, with $W > 18$). Exceptions were randomly chosen for each participant, with the constraint that each wind speed could be chosen only once. This meant that half of all possible wind speeds were presented with a 5:0 ratio of the two contexts, whereas the other half were presented with a 4:1 ratio. The first two sets of 18 training trials involved first one context and then the other, with the order of contexts counterbalanced across participants. Exceptions were absent during the first 36 trials to enhance the salience of the role of context (cf. Lewandowsky et al., 2000). The order of exceptions was randomized across the remaining 144 trials.

In the randomized-context condition, in contrast, context was randomly paired with stimulus value, so that each stimulus could occur in each context between one and five times. The two contexts were presented with equal frequency. In all other respects, the condition was identical to the systematic-context condition, including the grouping of contexts during the first 36 trials.

In the left-only condition, all stimuli were presented in the back-burning context. For 90% of the stimuli, wind speed was below 18, and for the remaining 10%, it was greater than 18. In the right-only condition, the reverse arrangement was presented, with 90% of all stimuli involving wind

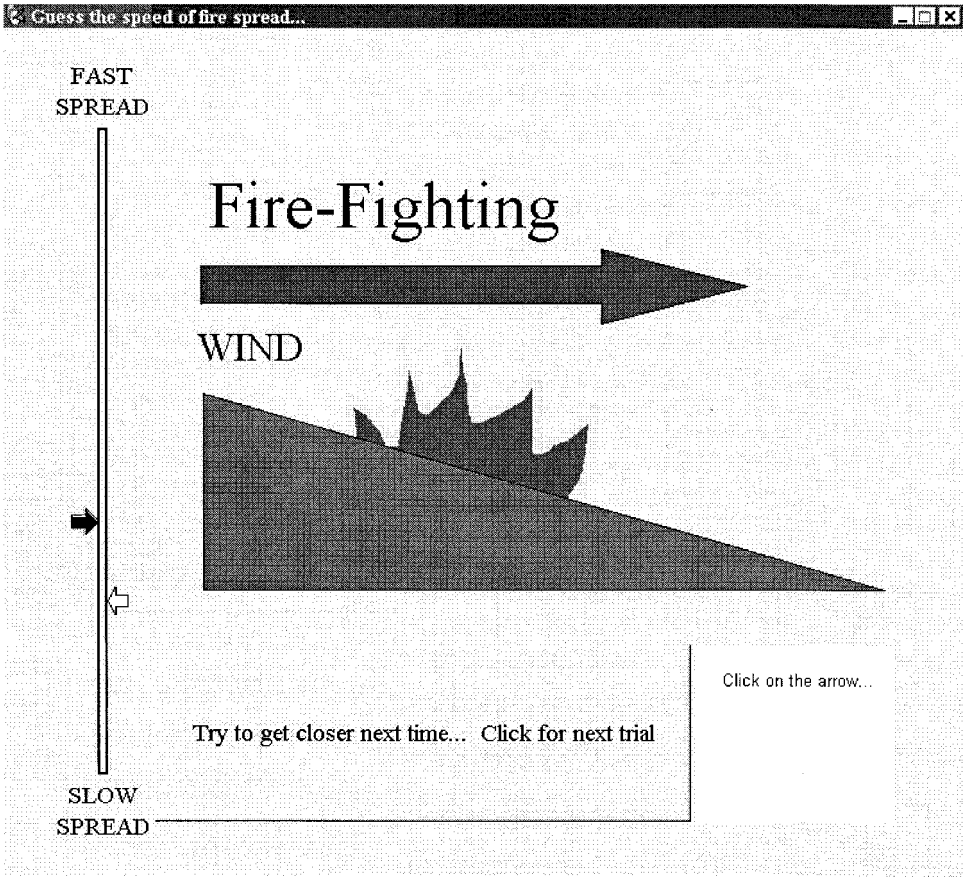

Figure 1. Image of the computer screen during a training trial after provision of feedback in Experiments 1 and 2. The horizontal arrow represents the stimulus magnitude; the vertical scale on the left was used by participants to indicate the response magnitude. The white arrow to the right of the scale represents the participant's response; the opposing black arrow represents the corrective feedback. Context is identified by the verbal label at the top (in this case, *Fire-Fighting*).

speeds greater than 18 and the remaining 10% involving wind speeds less than 18, all presented in the fire-fighting context. The exceptions occurred on randomly chosen trials after the first block.

After completion of the training trials, participants in all conditions completed the same transfer test. The transfer test involved predicting the fire speed for all 37 stimuli in both contexts, which yielded 74 transfer trials total. Participants were told that this transfer phase would include both new and old stimuli and that feedback would not be provided. The order of transfer trials was randomized separately for each participant. Experimental sessions lasted about 45 min.

Results and Discussion

Training Performance

Training performance was analyzed by considering the absolute deviations between the true function values and participants' responses. In the left-only and right-only conditions, deviations were averaged across each block of 18 contiguous trials, yielding a total of five observations per participant. In the systematic-context and randomized-context conditions, deviations were averaged separately for each context across blocks of 18 successive trials in a given context, yielding five observations per participant and context. The top panel in Figure 2 shows training performance across blocks for all conditions.

As can clearly be seen in Figure 2, performance improved considerably across training, which was confirmed by two separate analyses of variance (ANOVAs) for each pair of conditions. For the systematic-context and randomized-context conditions, the 2 × 2 × 5 (Condition × Context × Trial Block) between-within ANOVA yielded a main effect of trial block, $F(4, 96) = 22.43$, $MSE = 1.30$, $p < .0001$, and for the left-only and right-only conditions, the corresponding 2 × 5 (Condition × Trial Block) between-within ANOVA also yielded a main effect of trial block, $F(4, 92) = 11.47$, $MSE = 0.59$, $p < .0001$. The effect of condition did not approach significance in either analysis (both Fs < 1), although the effect of context was significant for the analysis of the systematic-context and randomized-context conditions, $F(1, 24) = 9.28$, $MSE = 0.53$, $p < .006$. This reflected the lower overall deviations in the back-burning context ($M = 4.16$) compared with those in the fire-fighting context ($M = 4.43$). No other effects approached significance, largest $F(4, 96) = 1.98$, $p = .104$, for the interaction between block and condition for the analysis of the systematic-context and randomized-context conditions.[3]

The obvious performance advantage of the left-only and right-only conditions over the other two was expected because, barring a few exceptions, the to-be-learned functions in these conditions were monotonic. It is well-known that monotonic functions are easier to learn than nonmonotonic functions (cf. Busemeyer et al., 1997), and hence the advantage of the left-only and right-only conditions is not considered further. Instead, for those two conditions, it is of interest to examine participants' performance on the exceptional stimuli that violated the monotonicity. The bottom panel of Figure 2 shows the average across participants of the responses given to the last training trial for each wind speed. Clearly, the low absolute deviations observed in the left-only and right-only conditions reflected performance on the majority of stimuli that conformed to the condition's prevalent monotonicity. Performance on the exceptional items, in contrast, was relatively poor. It must be noted, however, that each exceptional stimulus was presented only once during training. Any learning arising from the feedback associated with that single presentation will therefore be apparent in only the transfer data.

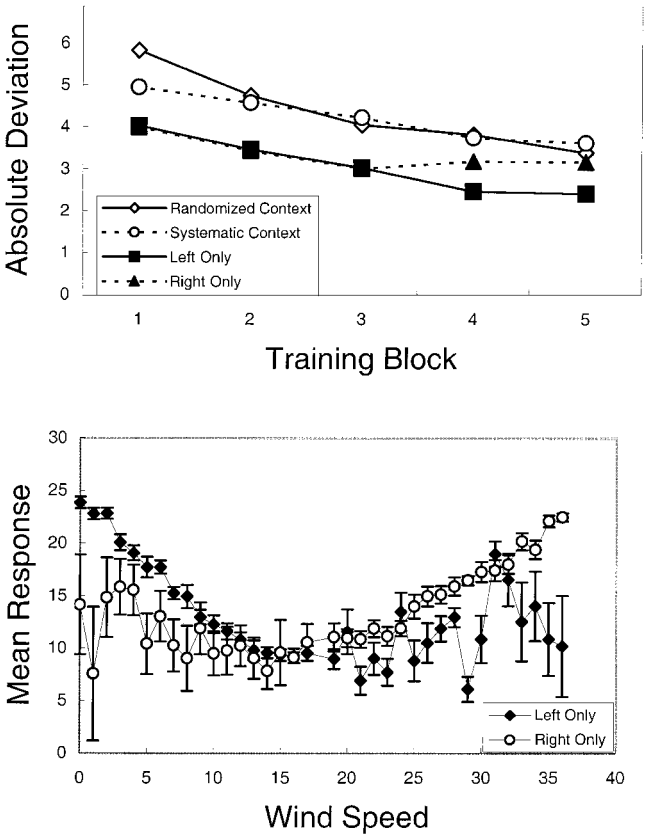

Figure 2. Training performance for all conditions across blocks of trials in Experiment 1. Top: Mean absolute deviations between responses and true values across all stimulus magnitudes. Bottom: Mean responses on the last training trial for each stimulus magnitude in the left-only and right-only conditions. Error bars indicate standard errors. Units are arbitrary.

Transfer Performance

Emphasis during the transfer analysis was on the effect of context across the four conditions. An overview of transfer performance is provided in Figure 3, which shows the responses for both contexts and all stimulus magnitudes, averaged across participants in the randomized-context and systematic-context conditions.

To facilitate visual assessment of performance, the panels also show the adjusted true function values. Virtually all participants overpredicted the function value at all stimulus magnitudes and in both contexts by a roughly constant amount. Because this constant bias was of little interest in the present context, for plotting

[3] Because this interaction was significant in some of the later experiments, and because it is of some theoretical interest, we explored it by comparing the systematic-context and randomized-context conditions using a between-groups t test for the first block of training. That difference also failed to reach conventional levels of significance, $t(24) = 1.70$, $p = .102$.

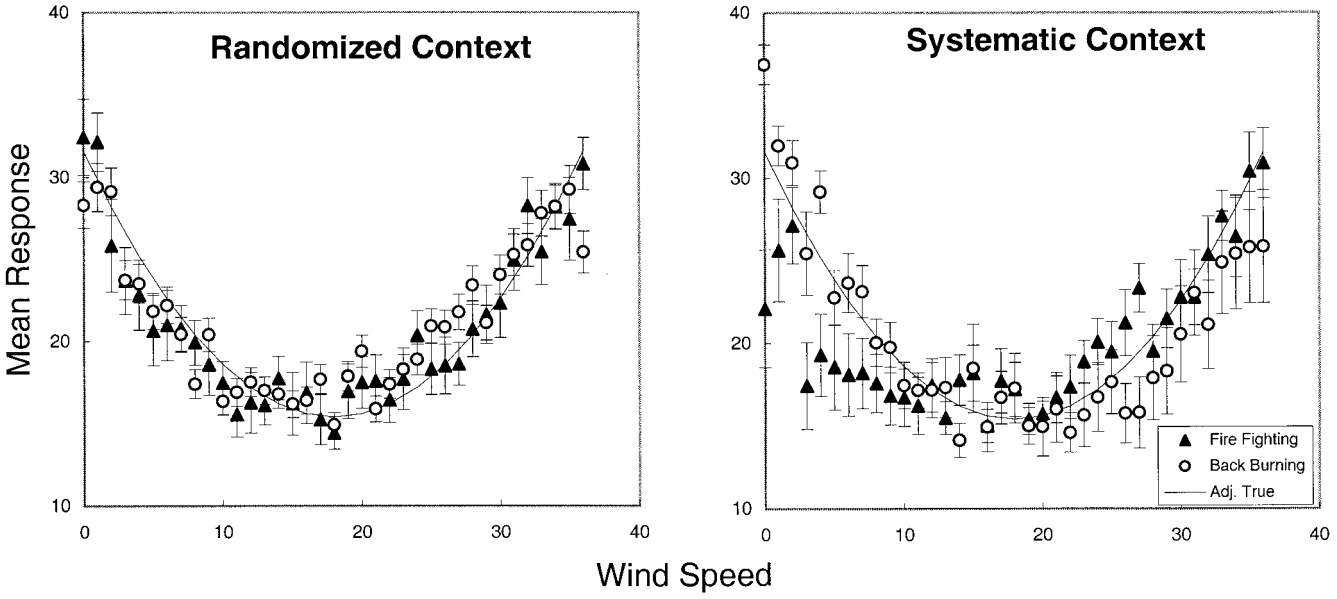

Figure 3. Transfer performance in Experiment 1 for the randomized-context (left) and the systematic-context (right) conditions. *Fire Fighting* and *Back Burning* refer to test context. The solid line in each panel represents adjusted (Adj.) true function values. Error bars indicate standard errors. Units are arbitrary.

purposes, each true function value was adjusted upward by a constant, computed by averaging the response bias across all transfer stimuli and all participants within a condition.

As can be seen in Figure 3, participants in the randomized-context condition learned the shape of the quadratic function quite accurately and their responses were relatively uniform across contexts. In the systematic-context condition, participants also learned the correct functional shape, although their responses appeared to be sensitive to context. In particular, extrapolations outside the predominant training context underestimated the adjusted true functional value, whereas those within the context overestimated the function. We discuss performance in the right-only and left-only conditions later.

Effects of context. The primary focus of the transfer analysis was on determining whether knowledge partitioning had occurred. This was achieved by computing the signed differences, for each stimulus magnitude and each participant, between transfer responses in the two contexts. If participants learned to integrate their knowledge across contexts, either because context was randomly paired with stimulus magnitude (in the randomized-context condition) or because people chose not to use it to gate performance even when it did have a second-order relationship with response magnitude (systematic-context condition), then those differences should reflect random fluctuation and should be roughly constant across all stimulus magnitudes. Conversely, if participants partitioned their knowledge by context when it was predictive of the functional relationship (systematic-context condition), then those differences should vary systematically with stimulus magnitude.

Figure 4 shows the signed differences averaged across participants in the randomized-context and systematic-context conditions as a function of stimulus magnitude. As can be seen in Figure 4, context had no effect at transfer when it was randomly paired with stimulus magnitude during training, as shown by the relatively uniform distribution of difference scores in the randomized-context condition. In contrast, context appeared to have an effect when it was predictive of the functional relationship, as shown by the clear relationship between difference scores and stimulus magnitudes for the systematic-context condition.

Statistical confirmation of this pattern was provided by a set of ANOVAs on the difference scores. First, an omnibus 4 × 37 (Condition × Stimulus Magnitude) between-within ANOVA revealed significant effects of condition, $F(3, 47) = 7.31$, $MSE = 346.43$, $p < .0001$, stimulus magnitude, $F(36, 1692) = 5.97$, $MSE = 76.94$, $p < .0001$, and, importantly, an interaction between those two variables, $F(108, 1692) = 2.28$, $MSE = 76.94$ $p < .0001$. A second between-within ANOVA included only the systematic-context and randomized-context conditions: the continued presence of a significant interaction, $F(36, 864) = 2.73$, $MSE = 69.70$, $p < .0001$, confirmed that the selective role of context involved the two conditions of greatest interest. Finally, the interaction was further explored by two within-subjects ANOVAs using stimulus magnitude as the only independent variable. For the randomized-context condition, stimulus magnitude had no measurable effect on the difference scores, $F(36, 432) < 1$, whereas for the systematic-context condition, that effect was clearly significant, $F(36, 432) = 3.25$, $MSE = 83.89$, $p < .0001$.

Together, the analyses confirm that knowledge partitioning occurred in the systematic-context condition. Specifically, the effect of stimulus magnitude on the difference scores indicates that participants accessed and used their knowledge on the basis of a second-order relationship between context and a to-be-learned outcome—that is, when it was predictive of the direction of the functional relationship. However, the analyses thus far do not specify the extent to which knowledge was partitioned in the systematic-context condition. Although partitioning was shown to

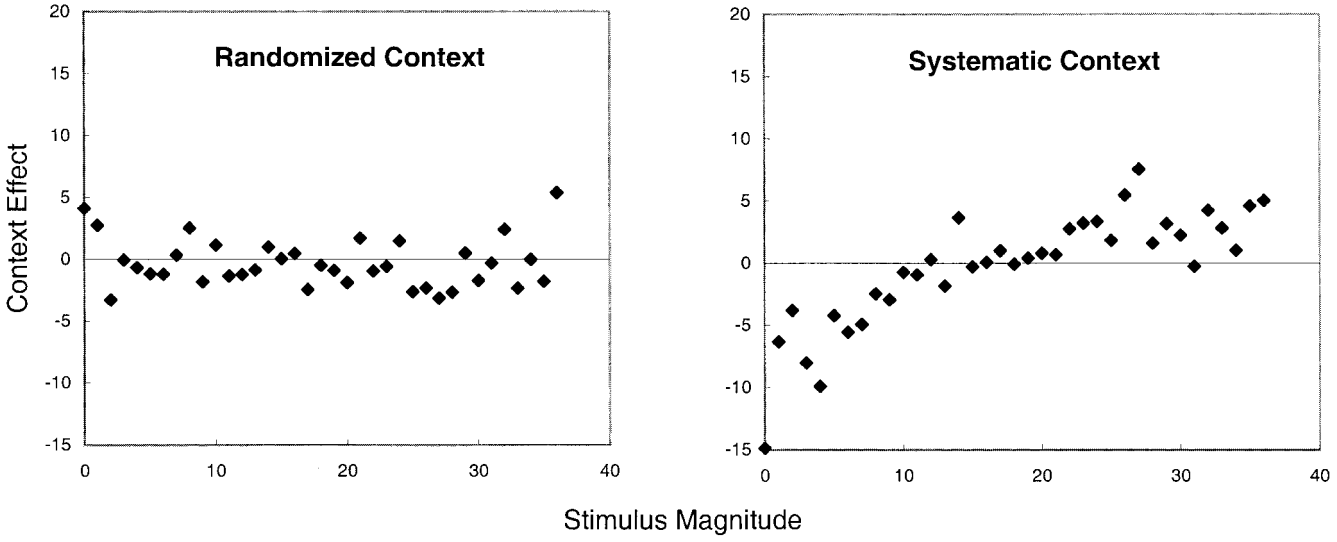

Figure 4. Average signed differences between test contexts for the randomized-context (left) and systematic-context (right) conditions in Experiment 1. Units are arbitrary.

be present, at first glance the numeric magnitude of the effect was arguably modest and seemingly incommensurate with the striking contradictions observed with experts (Lewandowsky & Kirsner, 2000; Schliemann & Carraher, 1993) and novices after training on a categorization task (Lewandowsky et al., 2000). It turns out that the interpretation of raw absolute magnitudes is misleading in the present experiment, and that the experimental design affords a better assessment of the extent of partitioning by considering performance in the left-only and right-only conditions.

Extent of knowledge partitioning. Consider the extreme case of complete partitioning in the systematic-context condition. Complete partitioning would imply that performance in one context was completely unaffected by information learned in the other context. In that case, irrespective of the absolute numeric magnitude of the context effect, transfer performance outside the prevalent training range for a given context should be identical to performance on the equivalent stimuli in the right-only or left-only conditions. Specifically, under complete partitioning, performance in the fire-fighting context at test should be identical for the systematic-context and the right-only conditions and, conversely, transfer in back burning should be identical for the systematic-context and the left-only conditions. In both instances, the comparison involves an identical prevalent training range within the context in question, the only difference being the presence of additional training in the alternative context in the systematic-context condition.

Figure 5 shows performance separated across panels by transfer context. In both panels, the systematic-context condition is compared with the appropriate control, involving the right-only condition for the fire-fighting context and the left-only condition the back-burning context, respectively. It is clear in Figure 5 that transfer performance differed considerably between contexts (i.e., panels), in line with the earlier analysis, but performance within a given context (panel) is highly similar across conditions.

This conclusion is supported statistically in several ways: First, considering the full set of 37 transfer stimuli, there are strong between-condition correlations within each context, $r = .88$ and $r = .91$, for fire fighting and back burning, respectively. This suggests that transfer performance is broadly identical within each context, irrespective of whether the other context was present at training. The impact of those correlations is best assessed by comparing them with the correlation between contexts in the systematic-context condition: Although that correlation is *within* subjects, it is far smaller ($r = .59$) than the within-context correlations that are computed *between* subjects.

Second, two separate between-within ANOVAs conducted for each context for all transfer stimuli failed to find a significant effect of condition in either the fire-fighting, $F(1, 23) = 2.03$, $MSE = 223.38$, $p > .10$, or the back-burning, $F(1, 24) < 1$, contexts. This again supports the assertion that transfer was unaffected by the presence of training in the other context. The ANOVAs did, however, reveal one effect involving condition that may have resulted from a subtle lack of commensurability between conditions for stimuli from within the training range: Condition interacted with wind in the fire-fighting context, $F(36, 828) = 1.49$, $MSE = 42.30$, $p < .05$. As shown in Figure 5, this interaction probably arose because responses differed between conditions for stimuli within the training range (i.e., $W > 18$), whereas responses seemed virtually identical in the extrapolation region (i.e., $W < 18$). In Figure 5, there is also a hint of a similar interaction for the back-burning context, with the suggestion of a difference between conditions in the training range (i.e., $W < 18$), although this interaction was not statistically supported.

We suggest that the interaction may have been caused by the differential spacing between repetitions of nonexceptional training items in the two conditions. Specifically, although the total number of training trials in a given context was equal between conditions, repetitions of nonexceptional items in the systematic-context condition were spaced twice as far apart, on average, as repetitions in the corresponding control condition (i.e., left-only or right-only). This was an inevitable consequence of the systematic-context condition containing the entire training sequence for the alternative

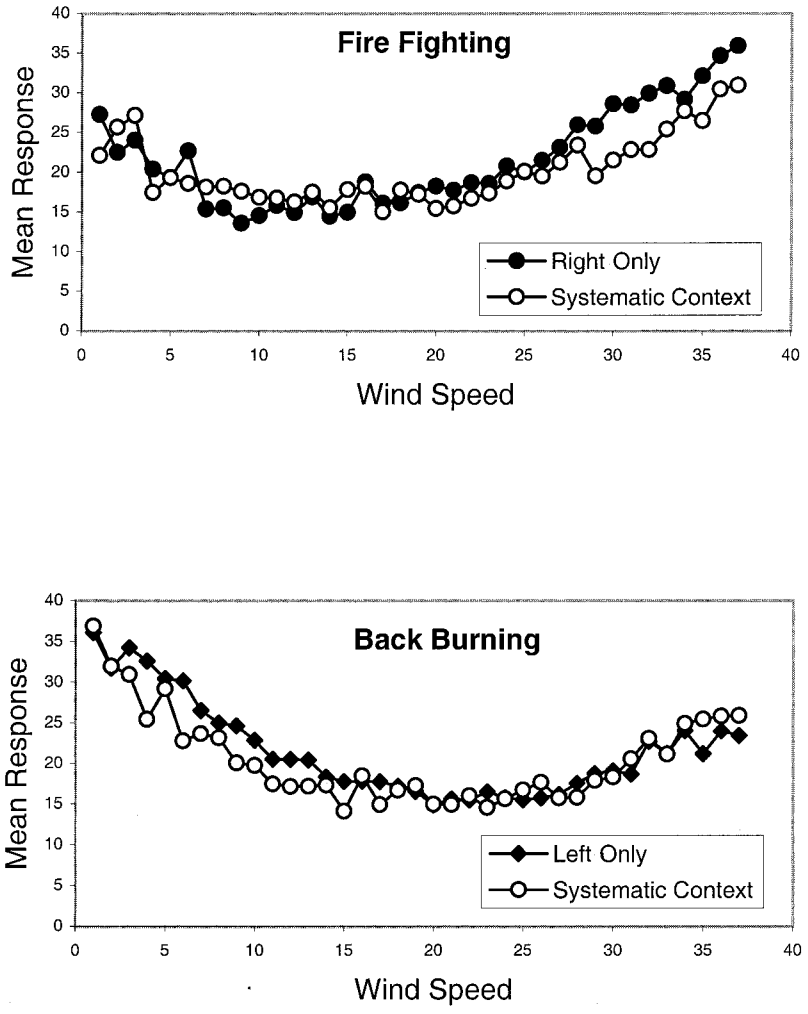

Figure 5. Comparison of performance in the systematic-context condition and the two control conditions in Experiment 1. In both panels, the systematic-context condition is compared with the appropriate control: the right-only condition for the fire-fighting context and the left-only condition for the back-burning context. Units are arbitrary.

context. The effects of spacing of repetitions on memory and learning are well-known and large (e.g., Toppino & Schneider, 1999), and this may have affected the within-training-range transfer responses. Fortunately, this was not an issue for the exceptional items, which in all conditions were presented only once and hence could not be differentially affected by spacing effects.

It follows that the most strictly commensurate comparison between conditions involves the transfer stimuli outside the training range—the extrapolation stimuli for which, as can be clearly seen in Figure 5, the conditions (in each panel) did not differ. In support, we repeated the separate between-within ANOVAs for each context on the subset of stimuli outside the prevalent training context (i.e., $W > 18$ for back burning and $W < 18$ for fire fighting). Condition had no effect in either analysis, nor did it interact with wind in either context (largest $F = 0.98$). (The F values for the main effects were 0.04 and 0.13 for back burning and fire fighting, respectively.) To guard against the possibility that this failure to reject the null hypothesis resulted from insufficient statistical power, we computed 95% confidence intervals for the differences between conditions for each context. Those intervals ranged from −4.50 to 5.48 and −3.22 to 4.61 for back burning and fire fighting, respectively. That they narrowly and symmetrically straddled zero supports our belief that for the stimuli whose training history was most commensurate, performance within each pair of conditions was identical.

We therefore conclude that transfer performance outside the prevalent training range was unaffected by learning in the alternative context, suggesting that knowledge partitioning in the systematic-context condition was complete or at least virtually complete. Because this conclusion is based on comparison with conditions that included no training in the alternative context, the arguably modest numeric magnitude of the effect has no bearing on the conclusion. At a theoretical level, we suggest that performance thus relied on one or the other parcel of knowledge, gated by context, with no discernible linkage or integration across parcels. This interpretation rests on the following facts: First, context

had no first-order relationship to the to-be-learned responses and thus could not have been integrated into a common weighted decision rule. Instead, because people were sensitive to its second-order relationship with response magnitude, context had to gate access to separate knowledge parcels. Second, because context-bound extrapolation was unaffected by the presence of training in the alternate context, the partitioning between parcels was complete.

The implications of complete partitioning can be clarified by considering the study by Aha and Goldstone (1992) mentioned at the outset. Aha and Goldstone found that people created a heterogeneous representation of a two-dimensional category space, in which training items were sampled from two separate clusters. In one cluster, the boundary between categories was horizontally oriented, and in the other, it was vertically oriented. Generalization performance relied on the appropriate boundary in the immediate vicinity of each cluster; however, the extent of that generalization was limited, because the two boundaries were globally incompatible (i.e., when extended further from their cluster, they dictated opposite classifications for the same test items). Thus, on the one hand, people's knowledge was arguably partitioned because a local boundary was applied for some categorization judgments, but on the other hand, the limited extent of generalization indicated that people's knowledge about each cluster was clearly influenced by the presence of the other set of training items. In the present experiment, in contrast, partitioning was virtually complete, because context-specific generalization was unaffected by the presence of competing knowledge.

Individual Differences

The analyses thus far considered only aggregate performance, without regard to individual differences. Individual differences are known to play an important role in function learning (N. Anderson, 1987; Carroll, 1963). For example, DeLosh et al. (1997) observed that nearly 20% of participants failed to accurately extrapolate a quadratic function on which they had been trained. In contrast, all participants were able to extrapolate a linear or (nearly linear) exponential function presented for training.

It follows that the knowledge partitioning observed in the systematic-context condition may have been due to the expected 20% of participants who were unable to learn a quadratic function and who may therefore have resorted to partitioning of the function into linear (or nearly linear) components to master the task at all. The corresponding subset of participants in the randomized-context condition would have found learning to be very difficult, because there was no independent predictor to permit partitioning.

This scenario would not negate our earlier conclusions about the existence and extent of knowledge partitioning. It would, however, suggest that partitioning is a strategy of last resort that people use only if they cannot otherwise master a complex task. As a first step toward exploring this possibility, we analyzed the individual differences in the randomized-context condition.

For each of the 13 participants, transfer responses in the two contexts were regressed separately onto the two predictors, W and W^2. This two-predictor model represents the correct parameterization of the to-be-learned function; thus, participants who learned the function can be identified by a high value of R^2 for this model. The mean R^2 across participants was .58 (range = .07–.86) and .56 (range = .02–.87) for the fire-fighting and back-burning contexts, respectively. Figure 6A shows transfer performance in both contexts for the participant who contributed the highest values of R^2.

Three participants (i.e., 23% of the total number in this condition) were clearly unable to learn the function in either or both contexts, with R^2 values at or below .10. Their responses are shown in Figures 6B–6D. The proportion of participants unable to learn a quadratic function is quite close to the value of 20% reported by DeLosh et al. (1997), and it suggests that there may be a similar number of participants in the other conditions who were likewise unable to learn a quadratic function. However, there is little point in examining individual differences in the systematic-context condition, because any departure from the correct quadratic form may be the result of knowledge partitioning. Knowledge partitioning, in turn, may arise for one of two reasons: The person may be unable to acquire a quadratic function and thus partitions out of necessity, or the person may strategically choose to partition the task despite having the ability to learn a quadratic function. The only way in which those two alternatives can be disentangled is by examining people's ability to learn quadratic functions before they are given the opportunity to use context to partition their knowledge. This was the purpose of Experiment 2.

Experiment 2

In Experiment 1, people's ability to learn a quadratic function could not be observed outside the experimental manipulation of interest. The second study was designed to ascertain people's ability to learn a quadratic function before context was manipulated. People participated in two sessions that ostensibly belonged to different and unrelated experiments. During the first session, participants learned a convex quadratic function without any manipulation of context. The purpose of this session was to identify participants who were, in principle, able to learn a nonmonotonic function in the time available. The second session involved a virtual replication of the systematic-context and randomized-context conditions of Experiment 1, using the same concave quadratic function and an identical context manipulation. The right-only and left-only conditions were not included in this study.

Method

Overview and Design

All participants took part in two sessions, a screening session followed by an experimental session, that were conducted on consecutive days. Participants were told that the two sessions belonged to two different studies that had nothing to do with each other. Accordingly, different experimenters tested participants in the two sessions and care was taken that the experimental software and written instructions had a different look and feel for each session.

In the screening session, all participants had to learn a convex quadratic function. The session involved 200 training trials, all of which included feedback. No experimental variables were manipulated and no context was specified when stimuli were presented.

The experimental session included only those participants who demonstrated during the screening session that they were able to learn a quadratic function. The experimental session was a virtual replication of the systematic-context and randomized-context conditions of Experiment 1. Participants were randomly assigned to one of the two conditions and then

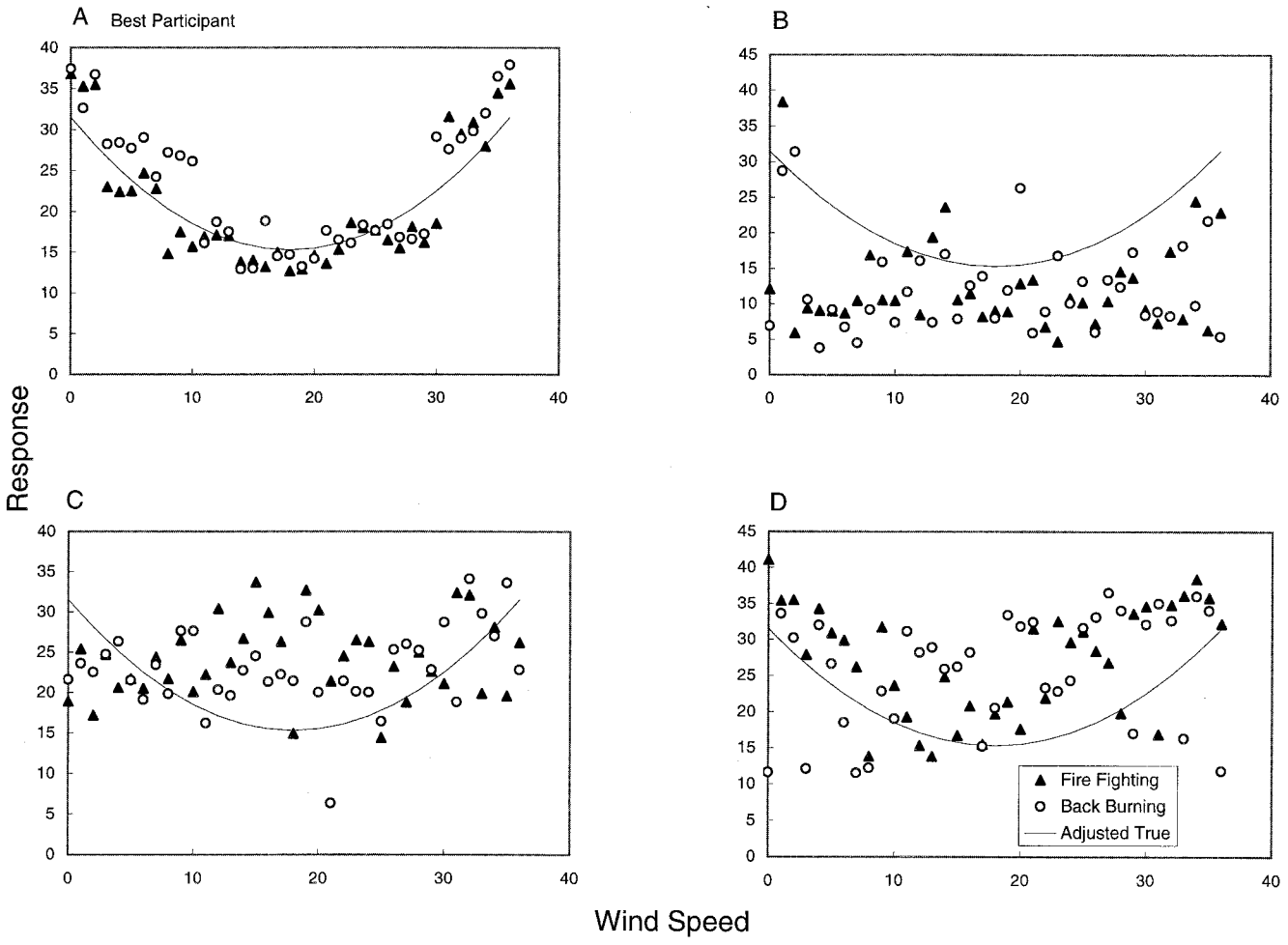

Figure 6. Transfer performance in both contexts for the participant (A) in the randomized-context condition of Experiment 1 who showed the best fit of the correct function to responses and the 3 participants (B–D) with the worst fits of the correct function to their responses ($R^2 \leq .10$). Units are arbitrary.

received 180 training trials followed by 74 feedback-free transfer trials. As in Experiment 1, all stimuli were presented in both contexts at transfer, and the assignment of context to stimuli during training varied between conditions.

Participants

All participants were undergraduates at the University of Western Australia who participated voluntarily. There were 49 participants in the screening session, three of whom failed to learn the quadratic function and thus did not contribute to analysis of the experimental session. A further 5 participants were unavailable for the experimental session, or their data were lost due to equipment failure, bringing the total number of participants in the experimental session to 41. Twenty-two participants were randomly assigned to the systematic-context condition, and the other 19 were assigned to the randomized-context condition.

Apparatus and Stimuli

IBM-compatible computers presented stimuli and collected responses in both sessions. For the screening session, the to-be-learned function was the convex quadratic equation $y = -0.00156 - 0.04x^2 + 4x$. Participants were told that x represented the dosage of a hypothetical psychotropic drug *Effemerol* and that they were to learn the associated severity (y) of psychotic symptoms. Training stimuli were created by sampling 36 values of x symmetrically and evenly around the midpoint, in the range 0–100.

Values of x and y were represented graphically by two horizontal slide bars without any incremental values or tick marks. The bottom bar represented the value of x, and the top bar represented the value of y. On each trial, participants moved the slide bar for y to the predicted correct level and then clicked on an *OK* button. A third horizontal slide bar then appeared between the other two and showed the correct value of y. Responses within 5 units of the correct response were additionally accompanied by the text message *Well done!* and other responses were accompanied by *Try to get closer next time*. Feedback remained visible until the participant initiated the next trial by clicking on the feedback text. Trials were separated by a 1-*s* blank period.

For the experimental session, stimuli were constructed in the same manner as in Experiment 1.

Procedure

Screening session. The training sequence consisted of 200 randomly sampled trials. Each stimulus was presented at least five times, and a

randomly chosen subset was presented six times to fill the sequence of 200 trials. The selection of stimuli and order of trials were randomized separately for each participant. Participants were instructed to complete the task as accurately as possible. The screening session took about 40 min to complete.

Experimental session. The procedure for the experimental session was identical to the systematic-context and randomized-context conditions of Experiment 1.

Results and Discussion

Screening Session

To identify people who were unable to learn a quadratic function, we analyzed performance across the 200 training trials by considering the absolute deviations between the true function values and participants' responses. Deviations were averaged across each block of 20 trials, yielding a total of 10 successive observations per participant. The means across participants declined monotonically from 25.71 during the first block to 9.68 on the last block, demonstrating that there was considerable learning across the 10 blocks of trials. This was statistically confirmed by a one-way within-subjects ANOVA, $F(9, 432) = 55.83$, $MSE = 22.92$, $p < .0001$.

To form an overall impression of participants' accuracy, we averaged all responses made for each stimulus magnitude during the last 20 training trials across replications (if any) and across participants. Those data are shown in Figure 7, with the standard deviations (not standard errors) associated with each mean response. Standard errors could not be computed because, owing to the randomization of the training sequence, the number of observations differed for each stimulus value and involved a variable combination of within- and between-participant variability. It is clear from Figure 7 that, as a group, the participants in the screening session were able to learn the function very accurately.

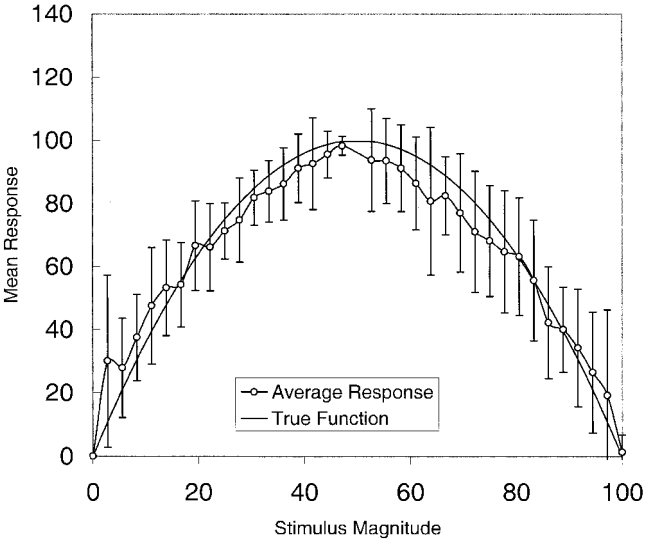

Figure 7. Mean responses during the last 20 training trials for all stimulus magnitudes during the screening session of Experiment 2. Error bars represent standard deviations (not standard errors). Units are arbitrary.

To identify participants who were unable to learn the function, we compared the absolute deviations between function values and responses during the final block of 20 trials across individuals. For three participants, the individual average deviation exceeded the group mean (9.68) by more than two standard deviations ($SD = 6.57$). In addition, each participant's responses during the final 20 training trials were entered into a separate regression analysis using the predictors x and x^2, akin to the individual differences analysis of the first experiment. Using a Bonferroni adjustment to the conventional significance level of .05, we found that the same 3 participants who were outliers on the deviation measure were also the ones for whom the correct regression model did not provide a statistically significant fit. Those participants were excluded from analysis of the experimental session. An additional participant similarly failed to show a statistically significant fit of the correct model, but because the individual's average deviation exceeded the group mean by little more than 1.5 standard deviations, that person's data were not excluded.[4]

Figure 8 shows the responses during the final training block for the excluded participants. For comparison purposes, Figure 8 also includes the responses of the participant with the smallest absolute deviation (1.65) and the best fit of the correct regression model ($R^2 = .995$), $F(2, 17) = 1{,}525.52$. It is clear from Figure 8 that the responses of the excluded participants, after 200 training trials, bore little if any resemblance to the true function.

Experimental Session

Training performance. As in Experiment 1, training performance was analyzed by averaging absolute deviations separately for each context across blocks of 18 successive trials in a given context. Figure 9 shows the training data.

It is clear from Figure 9 that the main findings of Experiment 1 were replicated. Performance improved considerably with training, and in both conditions, there appeared to be a slight advantage for the back-burning over the fire-fighting context. Both impressions were confirmed statistically by a 2 × 2 × 5 (Condition × Context × Trial Block) between-within ANOVA, which yielded a main effect of trial block, $F(4, 156) = 34.00$, $MSE = 1.91$, $p < .0001$, and a main effect of context, $F(1, 39) = 8.30$, $MSE = 1.33$, $p < .006$. The latter effect reflected the lower overall deviations in the back-burning context ($M = 3.63$) compared with the fire-fighting context ($M = 3.96$).

Unlike in Experiment 1, the analysis uncovered two significant interactions. The Condition × Trial Block interaction was significant, $F(4, 156) = 5.73$, $MSE = 1.91$, $p < .0001$, which reflected that performance in the randomized-context condition was poorer at the outset but at the end of training, exceeded that in the

[4] Responses by that same individual also could not be accounted for by the correct regression model in the experimental session.

A further 3 participants exceeded the mean absolute deviation computed across all participants by more than one standard deviation (but less than 2). It appeared inappropriate to exclude these additional participants from analysis of the experimental session as they did not unambiguously constitute outliers. Moreover, by fortuitous coincidence, all three were assigned to the randomized-context condition during the experimental session, which precludes these participants from contributing to a partitioning-as-a-last-resort effect.

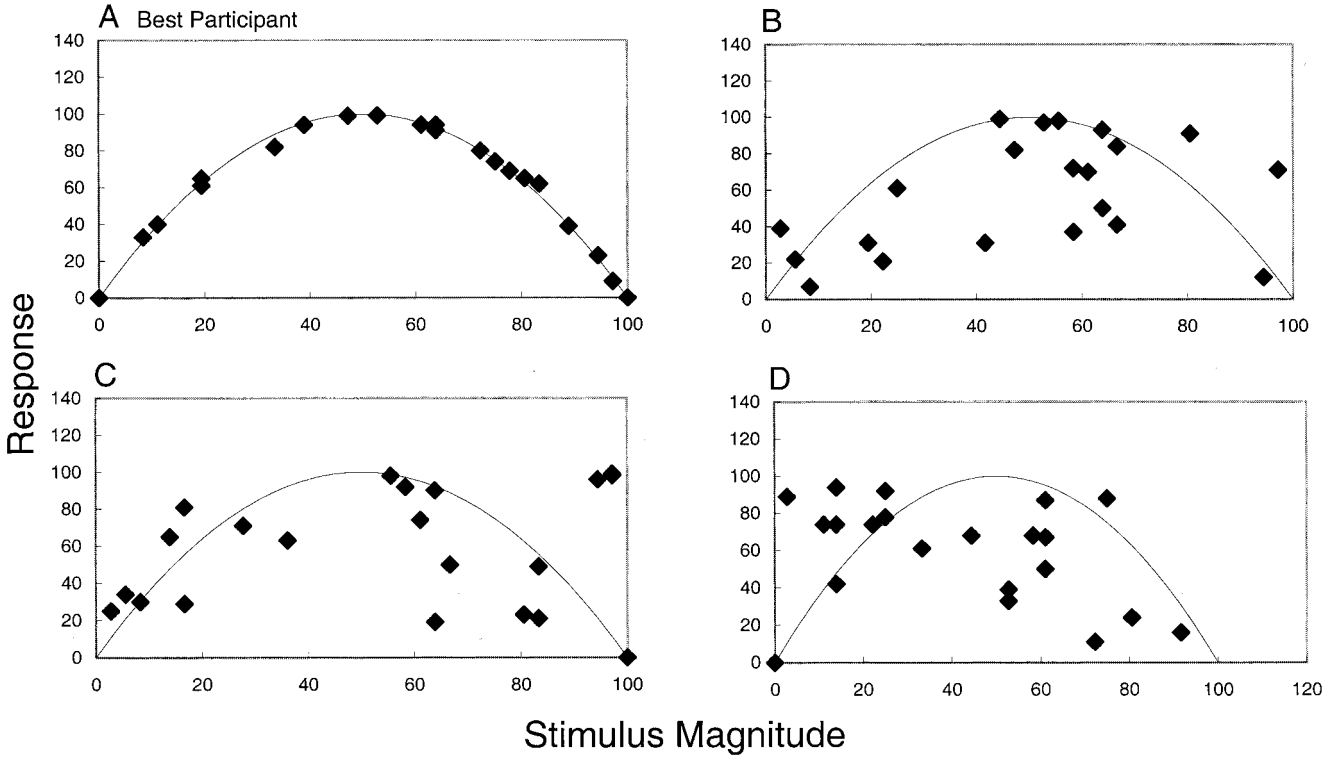

Figure 8. Responses during the final training block of the screening session in Experiment 2, for the best participant (A) and 3 participants (B–D) who were excluded from analysis of the experimental session because they failed to learn the quadratic function. Units are arbitrary.

systematic-context condition. The randomized advantage late in training likely reflected that people in the systematic-context condition ignored the exceptional stimuli outside the predominant training context, which prevented a further reduction of error. By the same token, ignoring of exceptions early in training permitted people in the systematic-context condition to rapidly learn monotonic functions, which, at the outset, engendered less error than people's attempts to learn a quadratic function in the randomized-context condition. We interpret this as being suggestive of the fact that immediate reduction in error is a principal reason why people with a proven ability to learn quadratic functions may choose to partition their knowledge. Once that choice has been enacted, it persists even if the early advantage turns into a cost later in training.

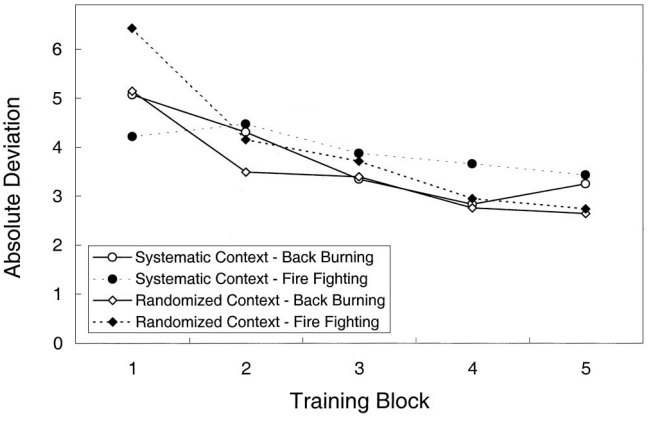

Figure 9. Training performance for all conditions across blocks of trials in the experimental session of Experiment 2. Units are arbitrary.

The three-way interaction involving all experimental variables was also significant, $F(4, 156) = 3.41$, $MSE = 1.75$, $p < .011$. We cannot offer an explanation for that interaction.

Comparing the magnitudes of the deviation scores between experiments, we found that performance throughout this experiment was better than that in Experiment 1. Specifically, whereas mean performance across the systematic-context and randomized-context conditions on the first block was 5.39 in Experiment 1, it was 5.17 in the present experiment. Performance on the final block was 3.50 and 3.04, respectively, in the two experiments. This performance advantage for Experiment 2 likely reflects the elimination of participants whom the screening session had identified as being unlikely to learn a quadratic function.

Stability of individual differences. An implicit assumption underlying the use of our screening session is that individual differences in the ability to learn a nonmonotonic function represent at least moderately stable attributes. For example, we expected a person who performed poorly during the first session to also do poorly on the next occasion, although the inflection of the function and the surface structure of the stimuli had been reversed.

To examine the validity of this expectation, we correlated participants' average deviations during the last training block of the

screening session with the corresponding score in the experimental session (taking the average across both contexts). To maximize commensurability between sessions, we considered only the randomized-context condition. To avoid truncation of the range of the variable under consideration, we entered all available participants into this analysis (including any who were identified as outliers). Finally, to correct an apparent nonlinearity in the relationship between sessions, we performed the analysis on log transformations of the average deviation scores. The resulting correlation was moderately large ($r = .40$, $N = 20$) and significant ($p = .04$, one-tailed).

We conclude that the ability to learn functions is at least moderately stable within a given individual. It follows that the elimination of participants who were unable to learn the function during the screening session reduces the likelihood that any partitioning we might have observed reflected a strategy of last resort by people who are otherwise incapable of learning a quadratic function.

Transfer performance. All participants responded to all stimuli twice, once in each context. Emphasis during analysis was again on the effects of context. Figure 10 shows the responses for both contexts and all stimulus magnitudes, averaged across participants in each condition.

The results again seem to replicate the main finding of Experiment 1. Participants in both conditions learned the correct functional shape, but in the systematic-context condition, responses were additionally sensitive to context. Thus, the partitioning of knowledge that was observed in Experiment 1 was replicated here, and as can be seen in Figure 10, the magnitude of the effect was even stronger than in the first study.

Paralleling the analysis of Experiment 1, we computed signed differences for each stimulus–participant between transfer responses in the two contexts. The crucial 2 × 37 (Condition × Stimulus Magnitude) between-within ANOVA on the signed differences revealed significant effects of stimulus magnitude, $F(36, 1404) = 3.47$, $MSE = 63.65$, $p < .0001$, and an important interaction between condition and stimulus magnitude, $F(36, 1404) = 2.89$, $MSE = 63.65$, $p < .0001$. Exploration of the interaction by two separate within-subjects ANOVAs using stimulus magnitude as the only independent variable revealed that the effect was limited to the systematic-context condition, $F(36, 756) = 4.69$, $MSE = 77.70$, $p < .0001$; for the randomized-context condition, $F(36, 648) < 1$.

The analysis confirms the presence of knowledge partitioning in the systematic-context condition. Given that all participants in this study were demonstrably able to learn a quadratic function, the results confirm that knowledge partitioning may be a fundamental aspect of learning in certain settings.

Individual differences. For each of the 19 participants in the randomized-context condition, transfer responses in the two contexts were again regressed separately onto the two predictors W and W^2. The mean R^2 across participants were .63 (range = .13–.84) and .57 (range = .06–.80) for the fire-fighting and back-burning contexts, respectively. Using a Bonferroni adjustment to the conventional significance level of .05, we identified 2 participants for whom the correct regression model did not provide a statistically significant fit of their responses. Both of those participants also performed relatively poorly in the screening session, although their deviation from the average in that session was insufficiently large to warrant their exclusion.

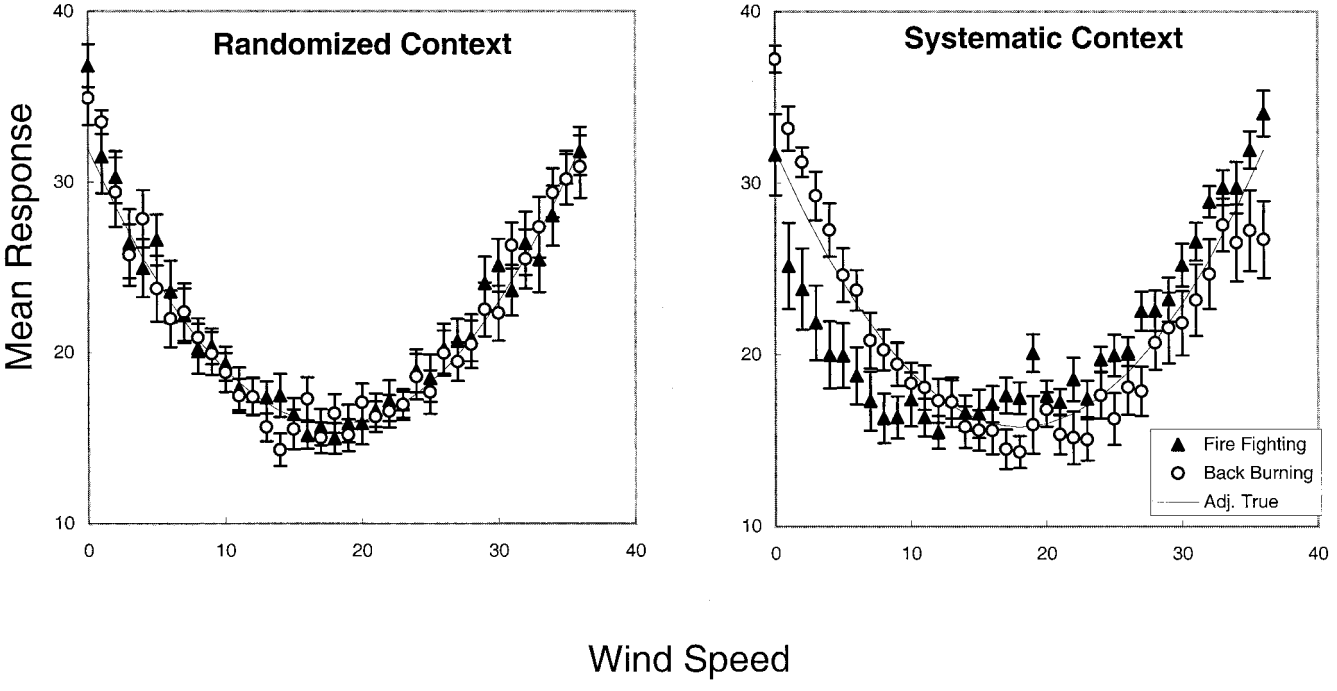

Figure 10. Transfer performance in the experimental session of Experiment 2 for the randomized-context (left) and the systematic-context (right) conditions. *Fire Fighting* and *Back Burning* refer to test context. The solid line in each panel represents adjusted (Adj.) true function values. Error bars indicate standard errors. Units are arbitrary.

Experiment 3

The first two experiments showed that people use a contextual cue that has no first-order relationship to a to-be-learned outcome but instead reliably predicts the relationship between two other variables (i.e., second-order relationship) to gate prediction performance. Having thus established the presence of knowledge partitioning under controlled conditions, we designed the remaining two experiments to explore the boundary conditions of partitioning. Specifically, we sought to investigate whether partitioning occurs when people are asked to learn a simple common function and the context variable is also predictive of the outcome—that is, we now introduced a first-order relationship between cue and response magnitude instead of the second-order relationship present in the first two experiments.

In the remaining two experiments, people learned a linear function. This function is known to be easier to learn and parametrically simpler than the quadratic functions used thus far (e.g., DeLosh et al., 1997). However, it is also known that ease of learning of linear functions differs with slope, with positive slopes being learned faster than negative slopes (e.g., Busemeyer et al., 1997). Accordingly, Experiments 3 and 4 used linear functions with a negative slope and a positive slope, respectively.

Analogous to the first two experiments, context was associated with different parts of the function in the systematic-context condition. However, unlike the preceding studies, the linearity of the function implied that the average response magnitude differed between contexts. In consequence, if people universally learn functions by relying on context—regardless of whether it is a first- or second-order predictor—generalization should again be context-specific. Indeed, on the basis of previous results (Lewandowsky et al., 2000) that showed people preferentially rely on a single imperfect predictor in a categorization task, partitioning might be expected to be particularly strong. Specifically, one might expect responses to cluster around the average response magnitude that was associated with a given context during training.

If, on the other hand, people rely on only context to simplify a complex task, for example, when context identifies components of a function that differ in slope or orientation, then generalization performance would not be expected to be context-specific, and people should learn the common linear function that transcends both contexts. This outcome would suggest that people seek to integrate their knowledge across contexts when it can be readily achieved, whereas (on the basis of the first two experiments) they partition knowledge when second-order context gates the components of a more complex function.

A third possibility is that if partitioning is a reaction to task difficulty, then the outcome would differ between experiments. Specifically, Experiment 3 may show partitioning but not Experiment 4, because negative functions are more difficult to learn than positive functions.

For Experiments 3 and 4, the experimental setting was changed to one involving the popularity of hypothetical radio show programs. Participants were asked to judge the popularity of radio shows after a programming change (Y) as a function of their initial popularity (X), for shows presented in either the morning or afternoon (context).

Method

Design

The design was similar to the experimental session of the preceding study. Participants were randomly assigned to one of two training conditions that differed with regard to the association between context and stimulus magnitude during training. In the systematic-context condition, context was predictive of stimulus magnitude, such that low values primarily occurred in one context (*morning*), whereas high values primarily occurred in the other (*afternoon*). Unlike the first two experiments, there was only a narrow band of stimulus magnitudes, in the middle of the range of possible values, that appeared in both contexts. Hence, the context manipulation segregated components of the function even more clearly than in the first two studies. The randomized-context condition was identical to that in the first two studies. During training in that condition, all stimulus magnitudes were equally likely to occur in either context.

Regardless of training condition, all participants were given a common transfer test in which all stimuli were presented twice, once in each context, thus yielding a $2 \times 2 \times 19$ (Training Condition $\times$ Test Context $\times$ Stimulus Magnitude) between–within-subjects design in the test phase.

During training, stimulus values were sampled from a window of 20%–80% of the total stimulus range. At transfer, stimulus values ranged from 5% to 95%, thus permitting examination of extrapolations outside the training range.

Participants

Twenty-nine undergraduates from the University of Western Australia participated voluntarily. Fifteen participants were randomly assigned to the systematic-context condition, and 14 were randomly assigned to the randomized-context condition.

Apparatus and Stimuli

The experiment was controlled by a IBM-compatible computer that presented stimuli and collected responses. The to-be-learned function was a linear equation relating initial ratings (R_0) to ratings after program changes (R_f), so that $R_f = 100 - R_0$ when both ratings are measured as proportions of the available scales. The results below are presented using that scale. Visually, from the participant's perspective, the stimulus (R_0) scale was 1.64 times as long as the response (R_f) scale. In consequence, in terms of absolute distance on the computer monitor, the to-be-learned function was $R_f = 61 - (0.61 \times R_0)$.

During training, 61 stimuli were sampled from the range 20–80, inclusive (for which 100 represented the maximum length of the stimulus scale). At transfer, the stimulus range was expanded but sampled less densely; the 19 test stimuli ranged from 5 to 95, inclusive, in steps of 5.

Each stimulus graphically represented initial rating and context. Initial rating was represented by a horizontal framed rectangle chart at the top of the screen, with the length of a brighter foreground region indicating the particular initial rating against a dark background of constant size. No numerical value for rating was provided. The color of the bar (green or red) coded the context in which the radio show should be considered. The assignment of color to context alternated between participants. Context was additionally identified by a textual label (i.e., *morning* or *afternoon*).

Participants were asked to predict the ratings of the radio show after the program change by using the mouse to move a slider on a vertically oriented scrollbar, which was presented on the right-hand side of the screen, away from the stimulus bar. The scale was labeled *low ratings* at the top and *high ratings* at the bottom, without any incremental values or tick marks.

During training, a response was immediately followed by an indication of the correct position on the vertical scale. The correct position was identified by an arrow that shared the context's color. Predictions deviating

by 5 or more units from the correct answer (approximately 1 cm on a 17-in. [43.2-cm] monitor) were followed by a warning tone and an admonition to try harder. Feedback had to be acknowledged by a mouse click. At transfer, feedback was absent. Stimuli were separated by a 2-s blank period.

Procedure

In the systematic-context condition, training stimuli in the range of 20–54 were presented in one context (morning), and stimuli in the range of 46–80 were presented in the alternate context (afternoon). This defined 70 possible stimuli, each of which was presented twice during the training sequence of 140 trials.

In the randomized-context condition, there were 61 possible stimulus magnitudes (from the set of 20–80, inclusive). Each stimulus was sampled at least once to generate the 122 training trials. On each trial, a context was randomly selected, and the two contexts were chosen with equal probability.

Regardless of training condition, all participants received each of the 19 transfer stimuli twice, once in each context. Transfer trials were presented in a different random order for each participant.

Results and Discussion

Training Performance

Training performance was analyzed in the same manner as in the first two studies: by computing the absolute deviations between responses and true function values and aggregating trials into blocks. Trials were aggregated into five blocks for each context, with approximately 13 trials in each block. (Owing to randomization of the experimental sequence, the number of trials per block varied slightly between and within participants.) Figure 11 shows training performance across blocks.

It is clear from Figure 11 that participants learned the linear function very rapidly and reached a high level of accuracy at the end of training. This was confirmed by the corresponding $2 \times 2 \times 5$ between-within ANOVA, which revealed a highly significant effect of training block, $F(4, 108) = 57.19$, $MSE = 12.25$, $p < .0001$. It is also apparent that participants in the systematic-context condition performed better throughout most of training, which is confirmed by the significant effect of training condition, $F(1, 27) = 5.88$, $MSE = 88.35$, $p < .05$. Because the randomized-context group made larger errors early in training but achieved a level similar to the systematic-context group by the end of the experiment, there was also a significant Block × Condition interaction, $F(4, 108) = 2.82$, $MSE = 12.25$, $p < .05$. To ascertain that both groups reached a comparable level of competence at the end of training, we conducted a t test comparing the two conditions for the last block of training, which was found to be nonsignificant, $t(27) = 1.01$. The 95% confidence interval associated with that difference ranged from −0.27 to 2.63.

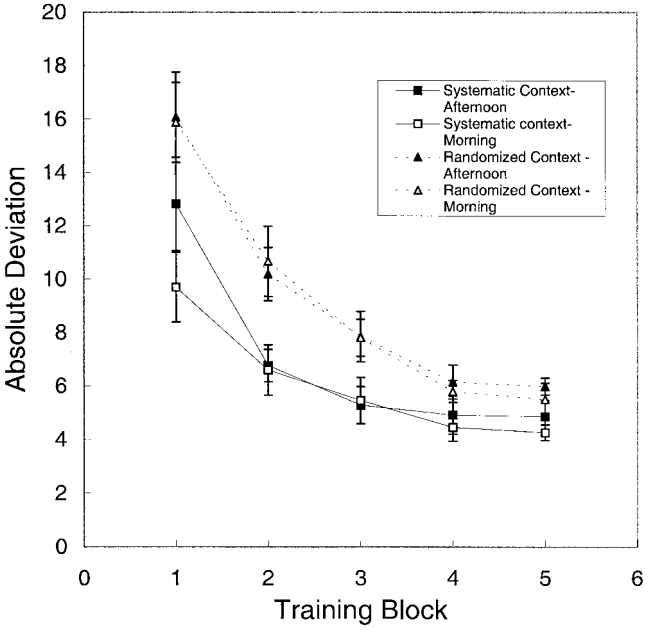

Figure 11. Training performance for all conditions across blocks of trials in Experiment 3. Error bars indicate standard errors. Units are arbitrary.

The interaction partially replicated that in Experiment 2, which also found an early advantage for the systematic-context condition over the randomized-context condition. As discussed previously, the immediate reduction in error may have been the primary motivation for people to partition their knowledge. No other effects in the ANOVA approached significance (largest $F = 1.69$).

It is interesting to consider the effect of context toward the end of training. Figure 12 shows mean responses in both conditions during the last two blocks of training. As expected, the randomized-context group showed no measurable effect of context—both response functions were indistinguishable. For the systematic-context group, in contrast, the two curves were clearly distinct. One way to explore this observation is to compare the responses of the two groups to stimuli within the overlap region (i.e., magnitudes of 46–54), for which both groups received identical training. The two-way between-within ANOVA on those stimuli involving condition and context showed no significant main effects, $F(1, 27) = 2.62$ and $F(1, 27) = 1.10$, for condition and context, respectively, but showed a significant interaction between the two, $F(1, 27) = 4.29$, $MSE = 19.83$, $p < .05$. The pattern of underlying means unequivocally identifies the difference between contexts in the systematic-context condition (47.79 vs. 51.36) as the source of the interaction, with the means in the randomized-context condition being virtually identical (51.82 vs. 50.54).

Thus, notwithstanding the context-invariant training regime for stimuli in the overlap region, responses differed between contexts for the systematic-context group. This context effect did not, however, affect overall accuracy of performance. As shown earlier, there was no significant difference between the two conditions for the last training block.

Transfer Performance

Transfer responses in both contexts are shown in Figure 13, averaged across participants in each condition.

The pattern in Figure 13 resembles that of the preceding experiments. As expected, no systematic difference between contexts is apparent for the randomized-context condition. For the systematic-context condition, in contrast, there is a clear difference between contexts across the entire range of stimuli that parallels that observed at the end of training (see Figure 12).

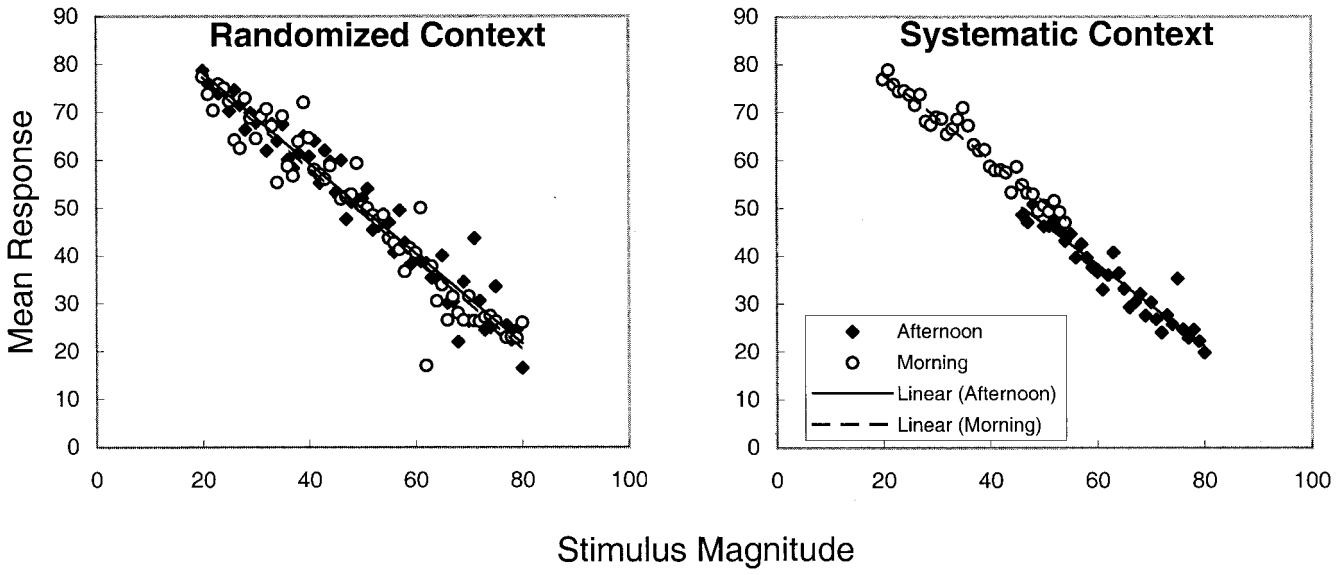

Figure 12. Mean responses in the randomized-context (left) and systematic-context (right) conditions during the last two blocks of training in Experiment 3. The lines represent linear regression models fit separately to mean responses in each context. Units are arbitrary.

As in the first two experiments, signed differences were computed for each stimulus magnitude and for each participant between transfer responses in the two contexts. The crucial 2 × 19 (Condition × Stimulus Magnitude) between-within ANOVA on those signed differences revealed a marginal effect of condition, $F(1, 27) = 3.32$, $MSE = 965.41$, $p = .079$, but no effects of stimulus magnitude, $F(18, 486) = 1.03$, and no interaction between the two experimental variables ($F < 1$).

Although the effect of condition was statistically marginal, it appeared sufficiently systematic in Figure 13 to warrant further exploration. This was done by considering the effect of context separately for each condition. Two separate analyses of the signed differences tested whether the context effect deviated from zero and revealed that the effect was present in the systematic-context condition, $F(1, 14) = 7.26$, $MSE = 769.50$, $p < .02$, but was absent in the randomized-context condition, $F(1, 13) = 1.57$. The

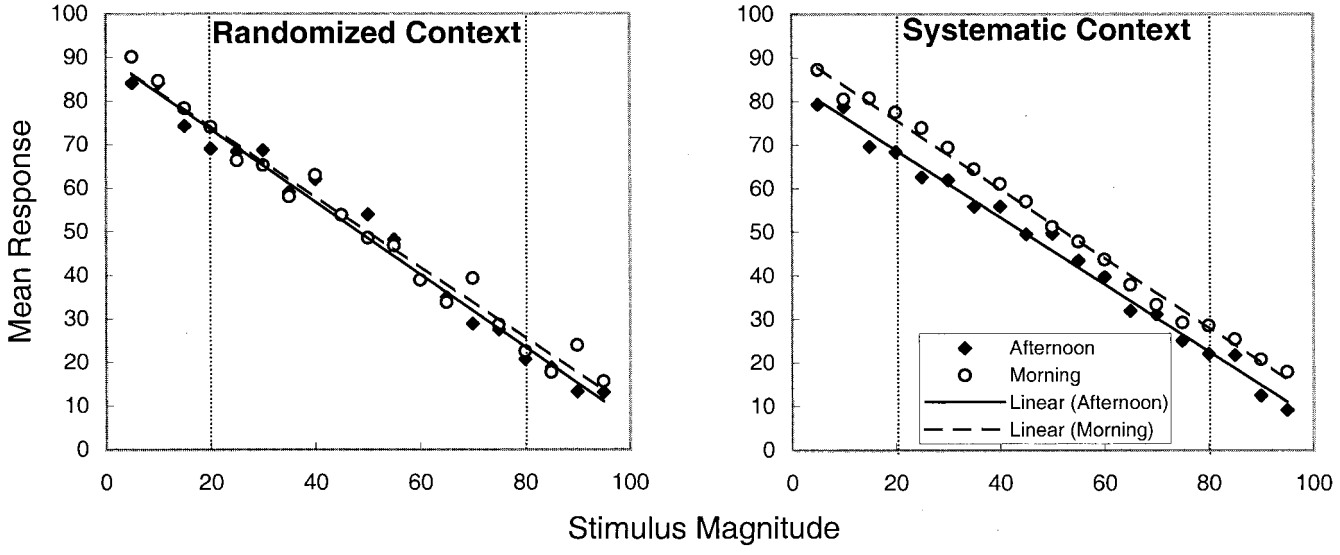

Figure 13. Transfer performance in Experiment 3 for the randomized-context (left) and the systematic-context (right) conditions. *Morning* and *Afternoon* refer to test context. The training region is between the vertical dotted lines. The solid and dashed lines represent linear regression models fit separately to mean responses in each context. Units are arbitrary.

effects were explored further by computing 95% confidence intervals of the context difference in each condition. For the systematic-context condition, the difference between contexts was 6.26, and the confidence interval ranged from 1.31 to 11.21; for the randomized-context condition, the difference was 1.43, and the associated interval was smaller and spanned zero, −1.02–3.88. Thus, the systematic-context condition engendered knowledge partitioning in a situation in which context had a first-order but no second-order relationship to the to-be-learned response.

An analysis of individual differences in the systematic-context condition revealed that the degree of partitioning differed across participants. This is illustrated in Figure 14, in which is shown the transfer performance of the 3 participants who exhibited the greatest visual extent of partitioning (B–D) and one representative individual who clearly and accurately learned a common function (A).

To rule out that the statistical effect of partitioning was due to the responses of a few individuals, we repeated the analysis of the systematic-context condition after the data of the 3 maximally partitioning participants in Figure 14 were excluded. Notwithstanding the removal of 20% of the observations that contributed most to the effect, the mean signed difference between contexts remained significantly different from zero, $F(1, 11) = 6.17$, $p < .03$. This suggests that partitioning was not confined to an identifiable minority of participants but represented a modal strategy.

The partitioning observed in Experiment 3 deserves to be contrasted with the findings of the preceding experiments. In the first two experiments, partitioning manifested itself by an increasingly greater context effect as the transfer responses extended further from the context-appropriate training range. In Experiment 3, in contrast, the context effect was uniform across the entire range of transfer stimuli. Across all three experiments, responses were

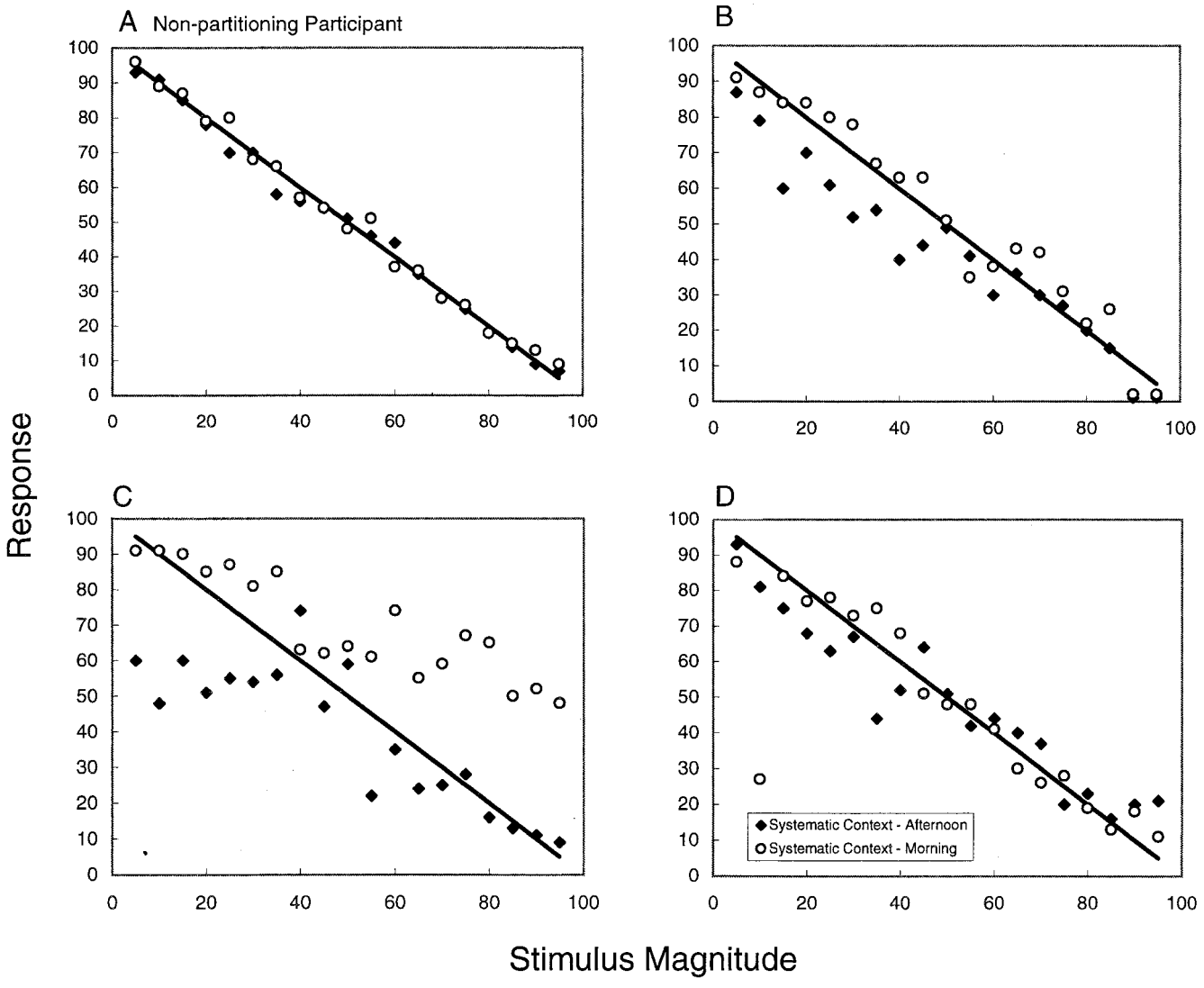

Figure 14. Illustrative individual differences during transfer in Experiment 3. A: Participant who learned a common function that transcended both contexts. B–D: Three participants who exhibited the greatest extent of partitioning by context. The solid lines indicate the true to-be-learned function. Units are arbitrary.

therefore in close agreement with the normative role of context, which predicted a functional relationship in the first two experiments and mean response (but not functional relationship) in Experiment 3.

Experiment 4

The final experiment was identical to Experiment 3, except that the linear function had a positive slope. Existing results show that an increasing linear function is not only learned most quickly (e.g., Busemeyer et al., 1997; DeLosh et al., 1997) but also conforms to people's a priori expectation of an unknown functional relationship (e.g., Sawyer, 1991). Experiment 4 therefore provided people with the easiest imaginable function learning environment. Should partitioning still occur under these circumstances, it would suggest that people are reluctant to ignore a context cue even when the common underlying function is maximally simple.

Method

Participants

Thirty undergraduates from the University of Western Australia participated voluntarily. An equal number of participants were randomly assigned to each of the two conditions. Data from 1 participant in the randomized-context condition was excluded because of equipment failure.

Apparatus, Stimuli, and Procedure

Apparatus and stimuli were the same as in Experiment 3, except that the slope of the function was reversed. In consequence, in terms of absolute distance on the computer monitor, the to-be-learned function was $R_f = 0.61 \times R_0$. In all other respects, this study was identical to Experiment 3.

Results and Discussion

Training Performance

Training performance was analyzed in the same manner as in Experiment 3 and is shown in Figure 15. It is clear from Figure 15 that participants learned the linear function very rapidly and reached a high level of accuracy at the end of training. This was confirmed by the corresponding $2 \times 2 \times 5$ between-within ANOVA, which revealed a highly significant effect of training block, $F(4, 108) = 14.90$, $MSE = 11.16$, $p < .0001$. No other effects approached significance (largest $F = 1.72$), confirming that neither context nor training condition affected speed of learning.

Comparison of the absolute level of performance at the end of training between Experiments 3 and 4 revealed that the positive linear function was learned to a much greater degree of accuracy than its negative counterpart. This is consonant with expectation and previous results.

Transfer Performance

Transfer responses in both contexts are shown in Figure 16, averaged across participants in each condition. The pattern in Figure 16 is very different from that observed in the preceding experiments: There is no evidence here that transfer performance was affected by context in the systematic-context condition. Paralleling the first three experiments, we again computed signed differences between transfer responses in the two contexts. The crucial 2×19 (Condition $\times$ Stimulus Magnitude) between-within ANOVA on those signed differences failed to reveal any significant main effects (largest $F = 2.79$ for the main effect of condition, $p = .11$). The interaction between condition and stimulus magnitude was also far from significant, $F(18, 486) < 1$.

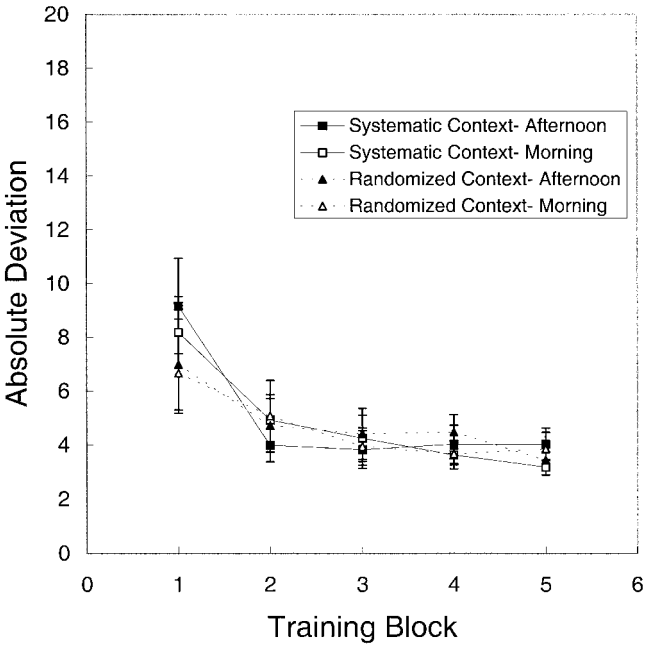

Figure 15. Training performance for all conditions across blocks of trials in Experiment 4. Error bars indicate standard errors. Units are arbitrary.

Because the *p*-value associated with the condition effect was moderately close to what might be considered noteworthy, the signed differences were again tested separately for each condition to determine whether they deviated from zero. For the randomized-context condition, that test was nonsignificant, $F(1, 13) = 1.75$. Most important, for the systematic-context condition, the corresponding $F(1, 14) < 1$. Together, these results clearly point to people having acquired a single linear function in both conditions. Thus, in contrast to the preceding studies, participants in the systematic-context condition did not consider context at all.

However, those conclusions rest on acceptance of several null hypotheses, such as the nonsignificant interaction between condition and stimulus magnitude and the nonsignificant deviations from zero of the signed differences. The conclusions would be strengthened considerably if it could be shown that the experiment had the statistical and methodological sensitivity to detect knowledge partitioning, had it been present. This is possible by considering extrapolation performance in both conditions.

Consider first the randomized-context condition. Figure 16 shows that extrapolations (stimulus values above 80 and below 20) depart slightly but consistently from the true function. In particular, participants' extrapolations above the training range (greater than 80) underestimated true magnitudes, whereas extrapolations below the training range (less than 20) resulted in overestimations. The signed deviations of responses from function values were

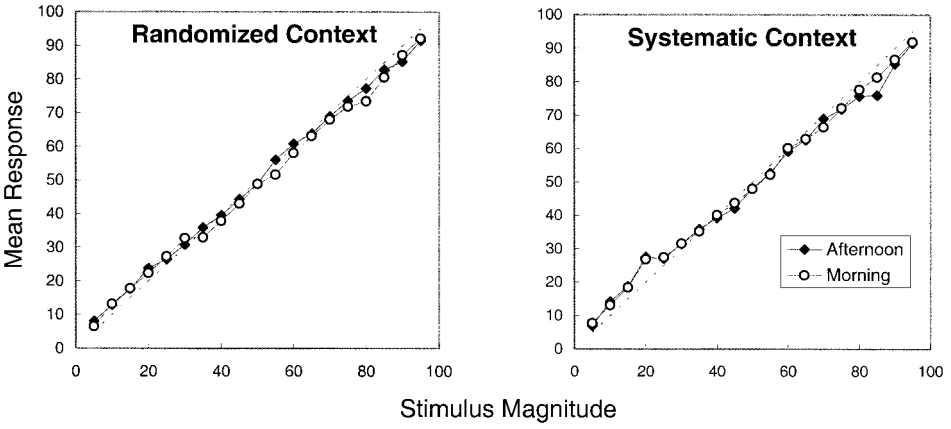

Figure 16. Transfer performance in Experiment 4 for the randomized-context (left) and the systematic-context (right) conditions. *Morning* and *Afternoon* refer to test context, and the true function is indicated by the dashed line. Units are arbitrary.

averaged separately within each extrapolation region (above and below the training range) for each participant and context. The top rows of Table 1 show the averages of those deviations across participants, together with the results of single-sample *t* tests that examined whether each average differed significantly from zero. Confirming the pattern in Figure 16, the results make it clear that the direction of deviation differed significantly with extrapolation region but not context.

Now consider the systematic-context condition. Paralleling the randomized-context condition, we found the extrapolations above the training range in the afternoon context and the extrapolations below that range in the morning context deviated significantly from the true values. Bearing in mind that afternoon and morning contexts were associated with training stimuli of high and low magnitudes, respectively, this result is not surprising. It confirms that extrapolation of a linear function beyond the range of trained stimuli is subject to systematic distortions.

Table 1
Average Signed Deviations of Responses From True Function Values (Means) During Extrapolation in Experiment 4

Context and extrapolation region	*M*	*t*	*p*
Randomized context condition			
Afternoon			
Above	**3.45**	**2.76**	**< .02**
Below	**−2.81**	**−2.91**	**< .02**
Morning			
Above	**3.36**	**3.14**	**< .008**
Below	**−2.48**	**−2.41**	**< .04**
Systematic context condition			
Afternoon			
Above	**5.78**	**3.91**	**< .002**
Below	0.96	< 1	> .37
Morning			
Above	1.62	1.42	> .17
Below	**−3.04**	**−4.40**	**< .0006**

Note. Significant deviations (as assessed by *t* tests with $df = 13$ for the randomized context and $df = 14$ for the systematic context) appear in boldface.

The remaining two deviations are of greater interest, because for them, *above* and *below* take on a context-specific meaning. Specifically, because the morning context was exclusively paired with low stimulus magnitudes during training, extrapolations above the training range in that context involved stimulus values of 55–65 rather than values above 80, as in the randomized-context condition. Conversely, extrapolations below the training range in the afternoon context referred to stimulus values of 35–45 rather than values below 20. In terms of distance from neighbors trained in the same context, the extent of extrapolation is identical to the above and below cases in the randomized-context condition. Nonetheless, these deviations from the true function values were much smaller and nonsignificant. Given that interpolations are known to be far more precise than extrapolations (e.g., DeLosh et al., 1997), the accurate responses in these extrapolation regions are best understood as *interpolations* that used responses learned in the other training context, as would be expected if participants acquired a single underlying linear function.

In summary, in contrast to the first three experiments, knowledge partitioning was absent in Experiment 4. Its absence is particularly noteworthy because, similar to Experiment 3, context had a first-order relationship to the required response and thus arguably provided more response-relevant information than in the first two experiments. Moreover, as shown in Table 1, Experiment 4 demonstrably had enough statistical power to detect deviations from the correct function values whenever extrapolation was required; we therefore conclude with some confidence that people acquired a common linear function.

General Discussion

Summary of Results

In three of the four experiments, people were found to tie new knowledge to the context in which it was acquired. This was revealed by the context-specific generalization observed when

context bore a first-order relationship to response magnitude (Experiment 3) as well as when it was normatively irrelevant but identified the predominant slope of the function (second-order instead of first-order relationship; Experiments 1 and 2). In all those cases, responses to the same stimulus differed between test contexts notwithstanding that a common underlying function had been presented for learning. Our favored explanation is that people partitioned their knowledge into independent parcels that may contain mutually contradictory information. The independence of parcels was underscored in Experiment 1, where context-specific extrapolation was unaffected by whether people had experienced extensive training in the other context.

Context was ignored only when people learned a linearly increasing function (Experiment 4). Linearly increasing functions are known to conform to people's preexisting expectations and are learned most rapidly (cf. Busemeyer et al., 1997; DeLosh et al., 1997). The selective absence of partitioning shows that knowledge will be integrated across contexts for only the simplest of all possible function learning problems.

Across all experiments, whenever partitioning was observed at transfer, it was associated with an advantage early in training compared with a control condition in which partitioning was prevented by randomly assigning contexts to stimuli. (However, in Experiment 1 this difference only approached significance.) In addition, there is a suggestion that partitioning persisted even when that early advantage turned into a cost later in training (Experiment 2).

Knowledge Partitioning: Empirical Status

Related Empirical Findings

We mentioned at the outset that there is considerable evidence for the coexistence of alternative strategies at all levels of skill acquisition and expertise. Although our results are related to these precedents, we now show that they also differ in two important respects.

Reder and Ritter (1992) and Schunn, Reder, Nhouyvanisvong, Richards, and Stroffolino (1997) repeatedly presented participants 2-digit $\times$ 2-digit multiplication problems (e.g., 43 $\times$ 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 two and three 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 coexist throughout the training sequence.

Prolonged coexistence of alternative knowledge has also been observed in 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 versus retrieving the answer from memory—may coexist 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, in press).

At a theoretical level, Lovett and Schunn (1999) presented a general model of choice during problem solving, known as *RCCL* (pronounced "ReCyCLe"), for which the central component is the availability of a set of alternative strategies to solve a common problem. According to RCCL, people choose a strategy for each trial on the basis of the past success of the available alternatives. In support, Lovett and Schunn found in two experiments that people choose among options on the basis of each strategy's past success rate. When no strategy was particularly successful, people repeatedly switched between them; conversely, people tended to persist with a strategy whenever it was successful.

In contrast to the present results and the knowledge partitioning framework, none of the foregoing studies presented evidence that these coexisting strategies could engender contradictory behaviors. That is, a problem such as 19 $\times$ 23 may be solvable by direct memory retrieval or by computation, and the two strategies may entail different completion times (e.g., Delaney et al., 1998; Schunn et al., 1997), but they both lead to the same correct answer. Indeed, should one strategy consistently give rise to errors that are avoided by use of an alternative, there is every reason to expect that people would rapidly abandon the less successful strategy (Lovett & Schunn, 1999). Moreover, none of the foregoing studies showed that performance of a strategy, once selected, was unaffected by knowledge embodied in another strategy; the present Experiment 1, in contrast, provided evidence of the independence of knowledge acquired in different contexts. We know of no other demonstrations of this independence and the promise of contradiction it entails, and thus, we consider this to be a unique feature of our results.

Knowledge Partitioning and Expertise

Experiments 1 and 2 used an analogue of the fire prediction task examined by Lewandowsky and Kirsner (2000), which had revealed context sensitivity in experts similar to that observed here with novices. The present results go beyond the earlier data in two important ways. First, in contrast to the expert study and its follow-up (Lewandowsky et al., 2000), the partitioning in Experiments 1 and 2 could not have reflected the incorporation of context into a single, multidimensional decision rule, because context had no first-order relationship to the required response. This eliminated one of the concerns cited at the outset: namely, that the partitioning observed previously might have been more apparent than real. Second, the nature of parcels here was experimentally controlled, thus eliminating the concerns about identifiability also cited at the outset.

Although these two aspects of the present methodology strengthen the viability of the knowledge partitioning framework, the earlier expert data still may not reflect partitioning. In the absence of experimental control over the complete acquisition history of expert knowledge, the relationship between the present findings and the partitioning sometimes observed in experts must remain circumstantial.

Nonetheless, it is tempting to accept a connection between those two manifestations of partitioning. In support, we can cite a competition between knowledge components that was observed across a longer time frame. This example differs from the earlier

work on multiplication and arithmetic because it revealed competition, rather than cooperation, among the coexisting strategies. Maloney and Siegler (1993) examined and compared the solution strategies applied to physics problems by undergraduate physics majors and nonmajors. Most relevant here is the finding that two thirds of all participants used three or more conceptually different strategies to solve the problems—although no more than two strategies were appropriate across all problems. The proportion of participants who used more than three strategies was the same for physics majors, who had on average enrolled in two university physics courses, and nonmajors, without any formal training in physics. These data suggest that (a) domain knowledge can encompass competing components, and (b) the competing components can coexist for considerable time periods. It thus remains possible that the partitioning in the present experiments is similar to that observed in experts (Lewandowsky & Kirsner, 2000). This possibility is further supported by the results of Mandl, Gruber, and Renkl (1992), who observed contradictions within domain-relevant declarative knowledge in a more complex microworld simulation of a factory. Their study, however, was limited by the extremely small sample size and because the participants could at best be considered advanced novices rather than experts.

If one accepts the possibility that expert knowledge may subsume the type of independent parcels created here in the laboratory, then one can identify at least one theoretical precedent in the expertise literature. This precedent involves the knowledge encapsulation framework mentioned at the outset when discussing the integrated nature of medical expertise (e.g., Boshuizen & Schmidt, 1992). In particular, knowledge partitioning shares the property of knowledge encapsulation that clusters of facts can be accessed through higher level concepts, but additionally assumes that components of knowledge are hidden from each other when encapsulated in separate parcels. This putative connection may merit further empirical exploration in the domain of medical expertise.

An empirical implication of knowledge partitioning involves the practice of knowledge elicitation. Knowledge elicitation refers to the process of explicating domain-specific expertise, usually for subsequent incorporation into a computerized expert system (e.g., Cooke, 1999). For example, medical practitioners might be asked verbally to report their diagnostic experience and knowledge, and that knowledge might then be embodied in an artificial intelligence application. On the partitioning view, the knowledge elicited from an expert would depend on the context in which it was elicited, and different contexts may give rise to different answers. A variant of this problem, known as the *differential access hypothesis,* has a long history of discussion among practitioners of knowledge elicitation (e.g., Hoffman, Shadbolt, Burton, & Klein, 1995), but it has until now been restricted mainly to consideration of the effects of different elicitation methods (e.g., verbal protocols vs. multidimensional scaling). The present results additionally suggest that identical elicitation methods presented in different contexts may elicit contradictory knowledge.

Knowledge Partitioning: Theoretical Status

As currently developed, knowledge partitioning is best understood as a broad framework for analyzing and understanding heterogeneity of knowledge and for guiding future empirical examinations of contradictory behaviors. Lovett and Schunn (1999, 2000) have recently underscored the crucial role of broad frameworks. Accordingly, we suggest that knowledge partitioning has successfully integrated findings from domains as diverse as expertise, categorization, and function learning under a common descriptive umbrella. The framework thus presents a platform for the future development of more precise but more narrowly applicable formal models that are required to verify the applicability of knowledge partitioning to a particular domain.

Although development of a formal model is beyond the scope of this article, we nonetheless foreshadow one possible approach for the function learning setting.

Theories of Function Learning

There are two primary classes of theories in function learning, one that claims the function concept is learned by memorizing exemplars and generalizing from those memories (Busemeyer et al., 1997; DeLosh et al., 1997), and another that claims the function concept is learned by parameter estimation on some intrinsic function, such as a high-order polynomial (e.g., Koh & Meyer, 1991). Busemeyer et al. (1997) convincingly argued that exemplar-based processes must be involved in function learning, with extrapolation from exemplars supported by an intrinsic linear function. The present experiments pose a serious challenge to this theory as formalized to date.

The formalism we refer to is EXAM (DeLosh et al., 1997), which extended an instance-based model of categorization (ALCOVE; Kruschke, 1992) to function learning by incorporating a linear extrapolation rule. Thus, EXAM learns a function by storing all stimulus–response pairs presented during training. Stimulus instances are arrayed along a single magnitude axis (X), and the extent to which each instance is activated depends on its numeric distance from the new stimulus.

When a novel transfer stimulus falls between two already-learned stimuli, the response is generated by taking the average of the learned responses to these previously encountered stimuli, weighted by their similarity to the new test item. When a novel stimulus falls beyond the training domain, then the response is generated by linearly extrapolating from the two old stimuli nearest the new one. As shown by DeLosh et al. (1997), these assumptions enable EXAM to accommodate a variety of results from function learning: in particular, the relative order of difficulty among linear, exponential, and quadratic functions. In all cases examined to date, EXAM provides a quantitatively more satisfactory account of the data than rule-based alternatives (DeLosh et al., 1997).

However, as it stands, EXAM is not designed to handle the presence of a binary context variable. Irrespective of any particular outcome, its application to the present experiments would require some modification to include a representation of context. We suggest one possible modification, but then show how even a modified version of EXAM is likely to encounter difficulties with the present results.

Given that EXAM is built on the ALCOVE model of categorization, which is known to be capable of representing multidimensional stimuli, it should be possible to augment EXAM accordingly. For example, if context were coded as {0,1} for the two alternatives, then EXAM could differentiate responses between contexts—The notions of *within* and *beyond* the training range that

apply to one-dimensional stimuli equally apply to two-dimensional ones. ALCOVE uses an additional construct: dimensional attention. The ability of ALCOVE (and by extension, EXAM) to differentiate between stimuli from different contexts depends on how much attention the context attracts. If attention to context were high, then EXAM would treat two stimuli with identical *X* values but different contexts very differently. If attention to context were low, then the same two stimuli would be treated more similarly to each other. Thus, the physical stimulus space of *X* is remapped into a psychological similarity space that stretches or shrinks as a function of attention.

The principal result of the first two experiments was the context specificity of extrapolation. For EXAM to reproduce this phenomenon, it must use the nearest stimuli from the appropriate context for its generalization, by separating the two sets of stimuli widely along the context dimension. To do so, the model must learn to pay attention to context. In principle, this would be possible, because ALCOVE can learn attention strengths for each dimension during training by focusing attention on the dimensions that are most relevant for the task at hand. This is done by evaluating the amount of error the network is making with its current attention strengths and then changing the strengths to reduce this error.

Nonetheless, we do not believe that this modification would enable EXAM to accommodate the outcome of all our experiments simultaneously. Attention learning should lead to a reduction in the strength of the context dimension in Experiments 1 and 2 (in which we found it to be strong experimentally) and a strengthening in Experiments 3 and 4 (in which we found it to be weak or absent). We now consider the reasons for this in turn.

In the design of Experiments 1 and 2, context was correlated with the magnitude of the stimulus but not the response. Hence, any attention paid to context would increase only error during training. Consider the case of a novel training item being presented during the course of the experiment. Because the item is novel, the model's response must be generated by generalizing (either interpolating or extrapolating) from other items in the psychological space. Because the true function being learned is smooth, the remembered responses to items of similar stimulus magnitude are always the best approximations to the correct response, regardless of their context. A response based on items of identical context but dissimilar magnitude will thus cause more error than a response based on items of similar magnitude but different contexts. Hence, any learning would diminish attention to context over time, in both the systematic-context and randomized-context conditions, quite at odds with the conditions identified earlier as necessary for EXAM to account for the results.

It follows that EXAM could only attempt to explain the results of Experiments 1 and 2 if attention was directed at context prior to learning and if learning then did not alter that attentional predisposition. (Of course, even with the assumption of static attention, it remains to be seen whether EXAM could handle the quantitative details of the results.) However, the possibility of static attention is placed in doubt by the unambiguous results of Experiment 4. In this experiment, any attention that was statically directed to context before learning would also engender context-specific extrapolation, in contrast with the data that showed participants clearly ignored context and learned a single underlying function. This result therefore eliminates the possibility that participants are always strongly and irreversibly attending to context—the only condition that might enable EXAM to handle the results of Experiments 1 and 2.

This conclusion is strengthened by the fact that in Experiment 3, under normatively identical conditions to Experiment 4 but with a decreasing linear function, participants were again shown to rely on context. It is unclear how the simultaneous presence and absence of attention with a linear function of different slopes can be accommodated by either static attention or a common attention-learning mechanism without ad hoc assumptions.[5]

In summary, EXAM could likely be modified to include a context dimension whose importance can be weighted by attention. This modification ought to be straightforward given ALCOVE's capability of multidimensional stimulus representations. Attention could in principle be either learned or statically predirected toward context. With either variant, it is difficult to see how EXAM could handle the results from all four experiments simultaneously: Whenever context should not attract any attention, because it is not predictive of an outcome, people were found to use it for partitioning. Conversely, whenever context was predictive of an outcome, people either ignored it or paid attention to it as determined by another characteristic of the to-be-learned function (namely, slope) that was of no relevance to attentional learning. We therefore suggest that potential theorizing must turn elsewhere to account for our results.

Toward a Model of Knowledge Partitioning in Function Learning

There is little doubt that expediency, the desire to acquire an efficient solution to a problem, is an important principle of human skill acquisition. For example, in categorization, people at least initially tend to choose the simplest of multiple possible classification rules (Lewandowsky et al., 2000; Medin, Altom, Edelson, & Freko, 1982; Nosofsky et al., 1994); in research on expertise, expediency has similarly been identified as a characteristic of expert knowledge (e.g., Abernethy, 1993; Hecht & Proffitt, 1995), and in a probabilistic prediction task, Edgell and Morissey (1987) observed the persistence of expedient initial learning even when a more complex configuration of cues became predictive later in learning. We suggest that this pervasive desire for expediency also underlies knowledge partitioning in function learning.

A critical benefit of expedient strategies is that, by definition, they are more easily acquired and hence reduce performance error more quickly than competing complex alternatives. In the present experiments, this was revealed by the lower mean absolute deviations in all systematic-context conditions that eventually led to partitioning (i.e., Experiments 1–3, although the effect was statistically unconvincing in Experiment 1). It follows that quick error reduction must be a hallmark of any model of partitioning in function learning.

We suggest that partitioning is best explained by a *mixture-of-experts* (MoE; Jacobs, Jordan, Nolan & Hinton, 1991) approach.

[5] DeLosh et al. (1997) enabled EXAM to learn increasing functions more quickly than decreasing ones by selectively presetting those weights to positive initial values that reflected a positive relationship between stimulus values and responses. This, however, has no bearing on the presumed attentional learning of context and is therefore unlikely to permit EXAM to account for the divergent pattern in Experiments 3 and 4.

An MoE model comprises multiple modules, each of which brings to bear partial expertise and competes to be chosen for governing the response on a given trial. At the outset, we cited evidence from categorization (Goldstone, 1996; Whittlesea et al., 1994) that demonstrated that people can selectively use alternative forms of knowledge. These ideas were embodied more formally in a recent MoE model of categorization, ATRIUM (Erickson & Kruschke, 1998), which combined an instance-based and rule-based approach. ATRIUM contained one module (or *expert*) for memorized instances and one module for each linear one-dimensional rule. Learning adjusted the relative importance of the two modules (instances vs. rules), the location of the decision rule on each of the categorization dimensions, and the associations of specific responses to each instance. ATRIUM implements an expedient approach to learning, because error is reduced very quickly when a simple rule can be learned that accommodates most responses.

A possible extension of this approach to function learning would likewise postulate several independent modules, each providing a linear mapping (of differing slopes and intercepts) between stimuli and responses. In this view, learning would consist of associating these various functions with the stimuli—either by finding the one function that works for all stimuli (as in Experiment 4) or by putting together a piecewise-linear approximation of the correct function.

In the systematic-context conditions, the presence of a context cue that is correlated with ranges of stimulus magnitude provides a convenient pointer to where the unknown function could be partitioned, thus facilitating learning and engendering a quick reduction of error. In Experiments 1 and 2, people might have associated positive and negative linear functions, respectively, with fire fighting and back burning. In Experiment 3, people might have associated each context with its own negative linear function. Once a function has been associated with a context cue, extrapolation is bound to that function irrespective of the other functions that may be associated with the same stimulus magnitude in other contexts.

In the randomized-context conditions, people may have instead learned a single function (e.g., in Experiments 3 and 4), or spliced together linear segments from different modules on the basis of stimulus magnitude rather than context. That is, when the common underlying function is linear, people may learn *Always use function F,* whereas when the function is quadratic, they need to learn *This stimulus falls within magnitude range X, so I will use function* F_X.

In this view, the difference in outcome between Experiment 4 and the other studies is due to people having a predisposition that favors a positive function. Because the function in Experiment 4 matches this disposition, alternatives do not gain any response strength, and hence, even in the systematic-context condition, no associations are formed between different values of context and different functions. In both conditions in that experiment, participants had no need to learn anything but *Use my preferred default function F.* We suggest that the future development of specific models along the lines just outlined can precisely instantiate the broad knowledge partitioning framework in several domains.

Conclusions

This article makes several empirical and theoretical contributions. Empirically, we showed that (a) in a function learning task, people create partitioned knowledge whenever possible, unless the to-be-learned function is of the simplest possible form. (b) People gate access to parcels of knowledge on the basis of context, (c) regardless of whether context normatively predicts response magnitudes. (d) Once knowledge is partitioned, use of information within each parcel is independent of knowledge present in other parcels. (e) Whenever partitioning is observed at transfer, it is associated with an advantage early in learning, although (f) it may persist even if it entails a performance cost later in learning.

At a theoretical level, our results (a) support and extend the few reports of conceptual competition observed in other domains, (b) appear incompatible with the preferred extant model of function learning, EXAM, and (c) point toward a mixture-of-experts approach.

References

Abernethy, B. (1993). Searching for minimal essential information for skilled perception and action. *Psychological Research, 55,* 131–138.

Aha, D. W., & Goldstone, R. L. (1992). Concept learning and flexible weighting. In J. K. Kruschke (Ed.), *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 534–539). Hillsdale, NJ: Erlbaum.

Anderson, J. R., & Fincham, J. M. (1994). Acquisition of procedural skills from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1322–1340.

Anderson, N. (1987). Function knowledge: Comment on Reed and Evans. *Journal of Experimental Psychology: General, 116,* 297–299.

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

Bédard, J., & Chi, M. T. H. (1992). Expertise. *Current Directions in Psychological Science, 1,* 135–139.

Boshuizen, H. P. A., & Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates, and novices. *Cognitive Science, 16,* 153–184.

Bott, L., & Heit, E. (2001, July). *Non-monotonic extrapolation in function learning.* Paper presented at the Third International Conference on Memory, Valencia, Spain.

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input–output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Knowledge concepts and categories* (pp. 405–437). Cambridge, MA: MIT Press.

Carraher, T. N., Carraher, D. W., & Schliemann, A. D. (1985). Mathematics in the streets and in schools. *British Journal of Developmental Psychology, 3,* 21–29.

Carroll, J. D. (1963). *Functional learning: The learning of continuous functional maps relating stimulus and response continua.* Princeton, NJ: Educational Testing Service. (ETS RB 63–26)

Charness, N., & Schultetus, R. S. (1999). Knowledge and expertise. In F. T. Durso, R. S. Nickerson, R. W. Schvaneveldt, S. T. Dumais, D. S. Lindsay, & M. T. H. Chi (Eds.), *Handbook of applied cognition* (pp. 57–81). Chichester, United Kingdom: Wiley.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152.

Cooke, N. J. (1999). Knowledge elicitation. In F. T. Durso, R. S. Nickerson, R. W. Schvaneveldt, S. T. Dumais, D. S. Lindsay, & M. T. H. Chi

(Eds.), *Handbook of applied cognition* (pp. 479–509). Chichester, United Kingdom: Wiley.

Cooke, N. J., & Breedin, S. D. (1994). Constructing naïve theories of motion on the fly. *Memory & Cognition, 22,* 474–493.

Delaney, P. F., Reder, L. M., Staszewski, J. J., & Ritter, F. E. (1998). The strategy-specific nature of improvement: The power law applies by strategy within task. *Psychological Science, 9,* 1–7.

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

Edgell, S. E., & Morissey, J. M. (1987). Delayed exposure to additional relevant information in nonmetric multiple-cue probability learning. *Organizational Behavior and Human Decision Processes, 40,* 22–38.

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

Ericsson, K. A. (1996). The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games* (pp. 1–50). Hillsdale, NJ: Erlbaum.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49,* 725–747.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology, 47,* 273–305.

Ericsson, K. A., Patel, V., & Kintsch, W. (2000). How experts' adaptations to representative task demands account for the expertise effect in memory recall: Comment on Vicente and Wang (1998). *Psychological Review, 107,* 578–592.

Even, R. (1998). Factors involved in linking representations of functions. *Journal of Mathematical Behavior, 17,* 105–121.

Fauconnier, G., & Turner, M. (1998). Conceptual integration networks. *Cognitive Science, 22,* 133–187.

Frensch, P. A., & Buchner, A. (1999). Domain-generality versus domain-specificity in cognition. In R. J. Sternberg (Ed.), *The nature of cognition* (pp. 137–172). Cambridge, MA: MIT Press.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12,* 306–355.

Glaser, R. (1996). Changing the agency for learning: Acquiring expert performance. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games* (pp. 303–311). Hillsdale, NJ: Erlbaum.

Gobet, F., & Simon, H. A. (1996a). Recall of rapidly presented random chess positions is a function of skill. *Psychonomic Bulletin & Review, 3,* 159–163.

Gobet, F., & Simon, H. A. (1996b). Templates in chess memory: A mechanism for recalling several boards. *Cognitive Psychology, 31,* 1–40.

Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). Strategic regulation of grain size in memory reporting. *Journal of Experimental Psychology: General, 131,* 73–95.

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition, 24,* 608–628.

Griffiths, T., & Kalish, M. (in press). A multidimensional scaling approach to mental multiplication. *Memory & Cognition*.

Hecht, H., & Proffitt, D. R. (1995). The price of expertise: Effects of experience on the water-level task. *Psychological Science, 6,* 90–95.

Hoffman, R. R., Shadbolt, N. R., Burton, A. M., & Klein, G. (1995). Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes, 62,* 129–158.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15,* 332–340.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 418–439.

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

Kimball, D., & Holyoak, K. (2000). Transfer and expertise. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 109–122). Oxford, United Kingdom: Oxford University Press.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus–response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 811–836.

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

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

Lewandowsky, S., Kalish, M., & Griffiths, T. L. (2000). Competing strategies in categorization: Expediency and resistance to knowledge restructuring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1666–1684.

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

Lewandowsky, S., Kirsner, K., & Bainbridge, J. V. (1989). Context effects in implicit memory: A sense-specific account. In S. Lewandowsky, J. C. Dunn, & K. Kirsner (Eds.), *Implicit memory: Theoretical issues* (pp. 185–198). Hillsdale, NJ: Erlbaum.

Light, L. L., & Carter-Sobell, L. (1970). Effects of changed semantic context on recognition memory. *Journal of Verbal Learning and Verbal Behavior, 9,* 1–11.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527.

Logan, G. D., & Klapp, S. T. (1991). Automatizing alphabet arithmetic: I. Is extended practice necessary to produce automaticity? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 179–195.

Lovett, M. C., & Schunn, C. D. (1999). Task representation, strategy variability, and base-rate neglect. *Journal of Experimental Psychology: General, 128,* 107–130.

Lovett, M. C., & Schunn, C. D. (2000). The importance of frameworks for directing empirical questions: Reply to Goodie and Fantino (2000). *Journal of Experimental Psychology: General, 129,* 453–456.

Maloney, D. P., & Siegler, R. S. (1993). Conceptual competition in physics learning. *International Journal of Science Education, 15,* 283–295.

Mandl, H., Gruber, H., & Renkl, A. (1992). Prozesse der Wissensanwendung beim komplexen Problem-Lösen in einer kooperativen Situation [Knowledge application processes during complex problem solving in a cooperative situation]. In F. Achtenhagen & E. G. John (Eds.), *Mehrdimensionale Lehr-Lern-Arrangements* (pp. 478–490). Wiesbaden, Germany: Betriebswirtschaftlicher Verlag Gabler.

Marchant, G., Robinson, J., Anderson, U., & Schadewald, M. (1991). Analogical transfer and expertise in legal reasoning. *Organisational behavior and human decision making, 48,* 272–290.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 37–50.

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

Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem solving, and induction. In R. J. Sternberg (Ed). *Advances in the psychology of human intelligence* (Vol. 5, pp. 189–223). Hillsdale, NJ: Erlbaum.

Moreau, C. P., Markman, A. B., & Lehmann, D. R. (2001). "What is it?" Categorization flexibility and consumers' responses to really new products. *Journal of Consumer Research, 27,* 489–498.

Müller, B. (1999). Use specificity of cognitive skills: Evidence for production rules? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 191–207.

Nosofsky, R. M. (1991). Typicality in logically defined categories: Exemplar-similar versus rule instantiation. *Memory & Cognition, 19,* 131–150.

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

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

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 510–520.

Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 398–415.

Nunes, T., Schliemann, A. D., & Carraher, D. W. (1993). *Street mathematics and school mathematics*. Cambridge, United Kingdom: Cambridge University Press.

Patel, V. L., Arocha, J. F., & Kaufman, D. R. (1999). Medical cognition. In F. T. Durso, R. S. Nickerson, R. W. Schvaneveldt, S. T. Dumais, D. S. Lindsay, & M. T. H. Chi (Eds.), *Handbook of applied cognition* (pp. 663–693). Chichester, United Kingdom: Wiley.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 435–451.

Reed, S. K., Dempster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 106–125.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General, 126,* 288–311.

Roth, E. M., & Shoben, E. J. (1983). The effect of context on the structure of categories. *Cognitive Psychology, 15,* 346–378.

Sawyer, J. E. (1991). Hypothesis sampling, construction, or adjustment: How are inferences about nonlinear monotonic contingencies developed? *Organizational Behavior and Human Decision Processes, 49,* 124–150.

Schliemann, A. D., & Carraher, D. W. (1993). Proportional reasoning in and out of school. In P. Light & G. Butterworth (Eds.), *Context and cognition: Ways of learning and knowing* (pp. 47–73). Hillsdale, NJ: Erlbaum.

Schmidt, H. G., & Boshuizen, H. P. A. (1993). On the origin of intermediate effects in clinical case recall. *Memory & Cognition, 21,* 338–351.

Schunn, C. D., Reder, L. M., Nhouyvanisvong, A., Richards, D. R., & Stroffolino, P. J. (1997). To calculate or not calculate: A source activation confusion (SAC) model of problem-familiarity's role in strategy selection. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1–27.

Shrager, J., & Siegler, R. S. (1998). SCADS: A model of children's strategy choices and strategy discoveries. *Psychological Science, 9,* 405–410.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116,* 250–264.

Spencer, R. M., & Weisberg, R. W. (1986). Context-dependent effects on analogical transfer. *Memory & Cognition, 14,* 442–449.

Tirosh, D., & Tsamir, P. (1996). The role of representations in students' intuitive thinking about infinity. *International Journal of Mathematics in Science and Technology, 27,* 33–40.

Toppino, T. C., & Schneider, M. A. (1999). The mix-up regarding mixed and unmixed lists in spacing-effect research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1071–1076.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352–373.

van de Wiel, M. W. J., Boshuizen, H. P. A., & Schmidt, H. G. (2000). Knowledge restructuring in expertise development: Evidence from pathophysiological representations of clinical cases by students and physicians. *European Journal of Cognitive Psychology, 12,* 323–355.

Vicente, K. J., & Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review, 105,* 33–57.

Whittlesea, B. W. A., Brooks, L. R., & Westcott, C. (1994). After the learning is over: Factors controlling the selective application of general and particular knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 259–274.

Received May 5, 2001
Revision received September 7, 2001
Accepted September 7, 2001 ■