Connection Science, Vol. 8, No. 2, 1996, 225–247

Dimensional Relevance Shifts in Category Learning

JOHN K. KRUSCHKE

(Received for publication April 1996)

A category learning experiment involving human participants compared the difficulties of four types of shift learning. Initial learning was of an exclusive-or (XOR) structure on two of three stimulus dimensions. One shift type was a reversal, a second shift was to a single previously relevant dimension, a third shift was to a single previously irrelevant dimension, and a fourth shift was to an XOR on one previously relevant dimension and one previously irrelevant dimension. Results showed that reversal shift was easiest, followed, in order, by shift to a single previously relevant dimension, shift to a single previously irrelevant dimension, and a shift to a new XOR. An extended version of the ALCOVE model, called AMBRY, qualitatively fits the data. The model incorporates two essential principles. First, internal category representations that can be quickly remapped to overt responses are important for accounting for the ease of reversal shift. Second, perseverating dimensional attention is important for accounting for the ease of shifting to a previously relevant dimension as opposed to a previously irrelevant dimension. It is suggested that any model of these effects will need to implement both of these principles.

KEYWORDS: Category learning, concept learning, discrimination learning, dimensional shift, relevance shift, reversal shift.

1. Introduction

In psychology, a traditional paradigm for studying learning and relearning examines how easily people accommodate various types of shifts in classifications or discriminations. For example, in one experiment the participant might initially have to learn that tall items belong in one category and short items belong in another category. The categorization is subsequently shifted, so that the participant must learn, for example, that green items belong in the first category and red items belong in the second category, regardless of height. Researchers have been interested in how learners respond to various types of shifts because of implications for theories of learning, generalization, attention and problem-solving.

There is a rich and extensive literature on the topic of shift learning, spanning several decades of research in the psychology of animal learning, human development, education and mental abnormality. To highlight this point, note that a literature review published nearly 30 years ago (Wolff, 1967) cited almost 200

J. K. Kruschke, Department of Psychology, Indiana University, Bloomington, IN 47405, USA. E-mail.

0954–0091/96/020225–23 $7.50 © 1996 Journals Oxford Ltd

publications on shift experiments in the subdomain of human learning alone, and a search of the PsycINFO bibliographic database returned 675 papers that included the phrase 'reversal shift' in their abstracts or keywords published in the years 1965–1995.

1.1. Previous Empirical Findings

What follows is a brief list of some of the basic effects observed in normal adult humans (cf. Zeaman & House, 1974). The effects are described in terms of 'classification' learning, in which a single stimulus is presented and the participant must classify it, but most of the effects were originally observed in 'discrimination' learning, in which two stimuli are presented and the participant must choose the stimulus that is an instance of the concept to be learned. Comparable effects are usually observed in both paradigms.

1.1.1. Reversal shift is easier than extra-dimensional shift. A 'reversal' shift simply reverses the correct classifications, so that what was in the first category is now in the second category, and vice versa. An 'extra-dimensional' shift changes the relevant dimension for classification. For example, the initial classification might be tall items in category 1 and short items in category 2, regardless of red or green color, and the shifted classification might be green items in category 1 and red items in category 2, regardless of height.

Extra-dimensional shift can be easier than reversal shift when novel values are introduced on the newly relevant dimension at the time of shift. For example, suppose the initial categorization depends on height, with (irrelevant) color varying between green and red. If, at the time of shift to color, the colors change to yellow vs blue, then the shift to color can be relatively easy.

1.1.2. Reversal shift is easier than non-reversal shift for items that span many dimensions. For example, suppose the stimuli to be classified are words, which span many orthographic and semantic dimensions. Suppose that participants learn initially to classify 10 unrelated words in one category and 10 other unrelated words in a second category. In a reversal shift, the correct category assignment of all the words is reversed. In this context, a reversal shift is also known as a 'pseudo-reversal' shift. In a 'non-reversal' shift, on the other hand, only half of words from each category are reversed. Despite more words changing their classification in pseudo-reversal shift than in non-reversal shift, pseudo-reversal shift is easier to learn.

1.1.3. After observing that a few items are reversed, normal adult humans tend to generalize that all items are reversed. This result comes from what is sometimes called the 'optional shift' design. Suppose, for example, that a participant has learned that tall items belong in the first category and short items belong in the second category. The participant subsequently sees short items classified into the first category (reversed relative to initial learning). When subsequently asked to classify tall items, the participant will tend to reverse their classification as well, despite having never observed any tall items explicitly reversed.

1.1.4. Intra-dimensional shifts are easier to learn than extra-dimensional shifts. An 'intra-dimensional' shift occurs when the same dimension is relevant in the shifted classification as in the initial classification, but the shifted classification is not a reversal, and the shifted classification depends on different values than the initial classification. For example, the initial classification could be green vs red, and the shifted classification could be yellow vs blue.

Intra-dimensional shifts are easier than extra-dimensional shifts when no novel dimensions are introduced at the time of shift, but not necessarily otherwise. For example, suppose that the initial classification depends on color, and all items have the same shape, but at the time of shift the items begin to vary in shape. In this case, shifting to a classification based on shape can be relatively easy.

1.1.5. Shifting from one 'compound' classification to another can sometimes be easier than shifting to a 'component' classification. A 'compound' classification is one that depends on more than one dimension. One case of a compound classification is the exclusive-or (XOR) structure: For example, an item is in the first category if it is green or square but not both. A 'component' classification is one that depends on just one dimension (or component), such as red vs green. Zeaman and House (1974, pp. 180–183) reported that participants who initially learned a compound discrimination could more easily learn a shift to another compound discrimination based on new values of the same relevant dimensions than a shift to a component discrimination based on new values of just one of the same relevant dimensions. This is a case of intra-dimensional compound shift, because the same dimensions were relevant. Barnes *et al.* (1978) reported that learning a compound discrimination is facilitated by previously learning a compound discrimination on other dimensions.

1.1.6. Summary. Normal adult humans are especially sensitive to reversal shifts, are especially sensitive to shifts to novel values or novel dimensions, are better at learning shifts that involve the same dimensions as were involved before the shift, and can sometimes learn compound-to-compound shifts better than compound-to-component shifts. There are, of course, many other interesting effects observed in studies of shift learning, such as the 'over-training reversal effect', in which reversal learning is faster when initial training is longer, and 'progressive improvement in repeated reversal learning' in which the speed of learning a reversal improves each time a reversal takes place.

1.2. Previous Theories

Theories of shift learning in the psychological literature have emphasized selective attention, 'mediating responses', and hypothesis or rule testing. To account for the advantage of intra-dimensional shift over extra-dimensional shift, some theorists suggested that participants learn to attend to specific dimensions, values on dimensions, or combinations of dimensions, and the learned attention persists or perseverates into the shift phase (e.g. Anderson *et al.*, 1973; Mackintosh, 1965; Zeaman & House, 1974). To account for the advantage of reversal shift, Kendler and Kendler (1962) posited the learning of internal, mediating responses between stimulus registration and overt response. The participant learns to associate various stimuli with particular mediating responses, which I prefer to call internal category representations, and the participant also learns to associate these mediating responses with overt classification responses. The mediating responses facilitate learning of reversal shift because the association between a mediating response and an overt response can be changed quickly without affecting the associations of stimuli to mediating responses.

Other theorists, such as Levine (1975), address these effects by suggesting that participants learn classifications by testing classification rules sampled from a hypothesis space. In Levine's (1975) theory, the hypothesis space has distinct domains and subdomains, such as the domains of all single-dimension rules, or all conjunctive rules, or all rules involving sequences of stimuli, etc. Subdomains include alternative rules involving the same dimension(s), such as rules related by reversal. Levine's 'transfer hypothesis' states that "When [the participant] receives a series of problems, he infers from the first *n* solutions the domain within the universe from which the $(n + 1)$th solution will be taken. He will start the $(n + 1)$th problem by sampling [classification hypotheses] from this domain . . . There is a tendency, furthermore, for [participants] to sample [hypotheses] within a subdomain before exploring other subdomains" (Levine, 1975, pp. 271, 276). The transfer hypothesis accounts for reversal-shift advantage because the reversed rule lies within the same subdomain as the original rule. If all compound rules lie within one domain, then the transfer hypothesis also suggests that shifting from a compound classification to another compound classification should be easier than shifting from a compound classification to a component classification.

There have been few attempts by connectionist modelers to address shift learning. Gluck and colleagues have applied two different models to the phenomenon of progressive improvement in learning of successive reversal shifts. Gluck *et al.* (1992) modified the configural-cue model of Gluck and Bower (1988) by including adaptive learning rates on each stimulus cue. Cues that are especially effective for error reduction are given larger learning rates. This can be construed as a form of selective attention to relevant cues. The modified model was able to learn reversal shifts much better than the original model, because learning rates on relevant cues grew larger. It remains to be seen, however, whether the modified model would learn reversal shifts faster than other types of shifts that also involve the same relevant dimensions, as is found in the experiment reported in this article. Gluck and Myers (1992) showed that their hippocampal model can learn successive reversals progressively more quickly. The hippocampal module of their model learns to discriminate stimuli that map to different categories better than stimuli that map to the same categories, and thereby implements a form of selective attention, although it is not restricted to particular psychological dimensions. The model also implements a form of mediating response, or internal category representation, by projecting the hippocampal representation into the cortical module. Gluck and Myers (1992) did not quantitatively fit their model to animal or human learning data, and it remains to be seen whether it can address the new data presented in this article.

1.3. Goals of This Article

In this article, I report results from an experiment that examines the relative difficulty of four types of shift after initial learning of a compound classification. One of the goals of the experiment was to discover whether shifting to a second compound classification would be easier than shifting to a component classification, and to discover whether shifting to a component classification that depended on a previously relevant dimension would be easier than shifting to a component classification that depended on a previously irrelevant dimension.

The results disconfirm some theories that posit attention to compound dimensions without attention to component dimensions. For example, a strong interpretation of the theory of Zeaman and House (1974) suggests that participants could learn to attend to the combination of color and shape because the combination is perfectly predictive of the classification, but learn to ignore the component dimensions of color or shape because they are individually irrelevant. One implication is that compound-to-compound shifts should be easier than compound-to-component shifts. The same prediction is made by a form of Levine's (1975) theory of hypothesis testing, in which compound rules involving the same dimensions lie within the same domain of the hypothesis space.

The results can be explained by two principles. First, people's attention to dimensions perseverates after the shift, so that it is easier to relearn categories with the same relevant dimensions. This principle agrees with some attentional theories mentioned above. Second, people's mental representation of the categories perseverates, so that it is easier to relearn a category-to-response change such as a reversal than to relearn category content. This principle agrees with the notion of mediating responses described earlier. The two principles are implemented in an extended version of the ALCOVE model, called AMBRY, which qualitatively fits the relative difficulty of the four shift types. In particular, it is also shown that AMBRY's dimensional perserveration alone cannot account for reversal shift advantage, for which mediating internal categories are needed. Thus, it is suggested that accounts of shift learning must incorporate both perseveration of dimensional attention and mediation of responses by internal category representations.

2. Experiment: Relative Difficulty of Four Types of Shift

Figure 1 shows the abstract structure of four types of shift presented to human learners. Stimuli varied on three binary-valued dimensions. The eight possible stimuli are represented in Figure 1 as circles at the corners of a cube, with each dimension of the cube corresponding to a dimension of variation in the stimuli. The color of each circle indicates its correct category assignment, with white circles indicating one category and dark circles indicating a second category. The cube at the left in Figure 1 shows the structure learned initially. This categorization is an XOR on the two horizontal dimensions (also called a biconditional rule), with the third (vertical) dimension being irrelevant. Relevant dimensions are indicated in Figure 1 by solid lines and irrelevant dimensions are drawn with dashed lines.

The right side of Figure 1 shows four different category structures learned subsequently. The top one is a 'reversal' shift, in which the categorization of every stimulus changes. The second one from the top is called the 'relevant' shift because the new categorization depends on a single dimension that was relevant in the initial XOR. The third shift categorization is called the 'irrelevant' shift because it depends on a single dimension that was irrelevant in the initial XOR. Finally, the fourth categorization is another XOR structure, involving one previously relevant dimension and one previously irrelevant dimension. It is called the 'compound' shift because the new categorization depends on two dimensions (and is not a reversal of the original categorization).

These four shift types were selected because they address some of the major

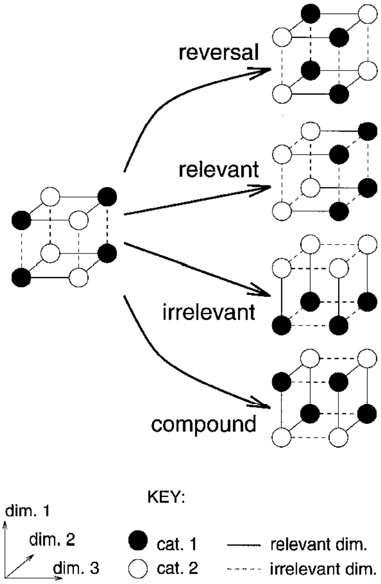

Figure 1. Abstract structure of four shifts used in the experiment with human learners. The cube on the left represents the structure of the categorization learned initially; the cubes on the right represent the four possible shifted categorizations. As indicated in the 'key', each dimension of variation is represented by an edge of the cube, each circle represents a stimulus such that the color of the circle indicates the correct category membership, and solid lines between circles denote relevant dimensions whereas dashed lines denote irrelevant dimensions.

phenomena in shift learning within a unified stimulus domain, all with the same initial learning. Different theories make different predictions regarding the relative difficulties of the four shifts. Theories that posit attention to compound dimensions, without attention to component dimensions, predict that the relevant and irrelevant shifts should be learned equally quickly, because neither of the component dimensions was initially attended to. Theories that posit a domain of compound rules, like one form of Levine's (1975) transfer hypothesis, predict that the compound shift will be learned faster than the relevant and irrelevant shifts, because the rule for the compound shift classification is in the same hypothesis domain as the initial classification. Alternatively, one might suppose that what matters is the number of dimensions that change their relevance, which predicts that reversal shift will be fastest (zero dimensions change relevance), relevant shift will be next fastest (one dimension changes relevance), compound shift will be next fastest (two dimensions change relevance) and irrelevant will be the slowest (all three dimensions change relevance). Yet another possibility is that what matters is the number of exemplars that change their categorization. This predicts that the relevant, irrelevant and compound shifts will be equally difficult because they all change the classification of four exemplars. Finally, as a sort of null hypothesis, one might suppose that all that matters is the structure of the shift classification and not its relationship to initial learning. In this case, the reversal and compound shifts should be equally difficult, because they are both XOR structures, and the relevant and irrelevant shifts should be equally difficult,

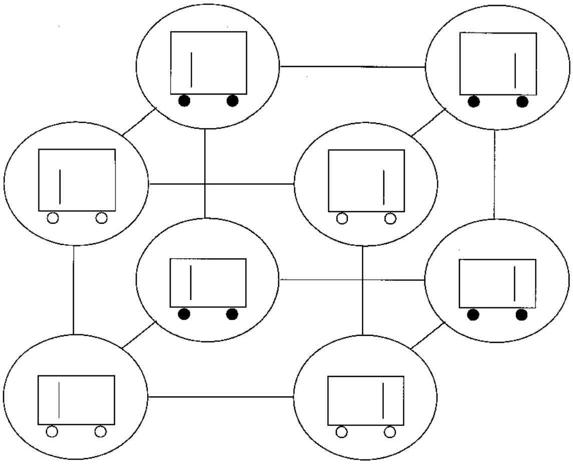

Figure 2. The eight stimuli used in the experiment with human learners. The assignment of these physical dimensions to the abstract dimensions of Figure 1 was counter-balanced across participants. (The large cube and ovals indicate dimensional relationships and are not part of the stimuli.)

because they both have a single relevant dimension. As will be seen, the results disconfirm all of these predictions.

2.1. Method

2.1.1. Stimuli. The three binary-valued dimensions were instantiated in schematic pictures of freight train box cars, illustrated in Figure 2. The eight stimuli are arranged in a cube in Figure 2, to emphasize the three binary-valued dimensions: height of box car (tall or short), position of the 'door' (left or right) and color of wheels (filled or blank). I chose these stimuli because I could safely assume the three dimensions were psychologically separable, based on previous work with rectangles and interior segments (Kruschke, 1993a). The three dimensions may have had different salience, however, and therefore were counterbalanced across participants, as described below.

2.1.2. Procedure. Training on the initial XOR structure lasted for 176 trials, i.e. 22 blocks of eight exemplars. Without any pause or announcement, the shift structure was introduced on trial 177, and training on the new structure continued for 80 trials, i.e. 10 blocks of the eight exemplars, for a grand total of 256 training trials. There was one break during the experiment, which occurred after trial 128 (halfway through the experiment), during the initial training.

Each participant saw a different random order of stimuli, with the constraint that each of the eight exemplars was shown once in each successive block of eight trials. Assignment of physical dimensions (height, door location, wheel color, category label) to abstract structural dimensions was counter-balanced across participants in order to counteract any differences in dimensional salience. For the initial XOR structure, there are three distinct mappings of physical dimensions to abstract dimensions, and two assignments of category labels. For the reveral shift structure, the realization of the initial XOR completely determines the reversal structure, so there are a total of six realizations of the reversal shift structure. For the relevant shift structure, there is a choice of two previously relevant dimensions, yielding 12 realizations. For the irrelevant shift structures, there is only one previously irrelevant dimension to select, so there are six realizations. For the compound shift structure, one of two previously relevant dimensions could be made irrelevant, yielding a total of 12 realizations. For each of the latter three shift structures, there are two possible assignments of category labels, but these were not independently varied relative to the random assignment of category labels in the first phase, in order to reduce the size of the design. Thus, in the reversal and irrelevant shift structures, each realization was run on 10 participants, and in the relevant and compound shift structures, each realization was run on five participants, making a total of 60 participants in each group.

Instructions were presented to the participant on the computer screen and were read aloud by the experimenter. The full text of the instructions is given in Appendix A. The instructions indicated that there were eight freight train box cars that varied on three dimensions of height, door position and wheel color, and that these three properties completely determined the route taken by the car. Participants were also told that there might be a time when the routes taken by the cars change, but that it would happen only rarely, and probably only after they had learned the routes perfectly. This latter instruction was given because it was found in pilot experiments that some participants would spontaneously interrupt the experiment when they noticed the route change, to warn the experimenter that something had gone wrong with the computer.

Stimuli were presented with a PC-type computer using VGA resolution, as white lines against a black background. Viewing distance was about 0.9 m, and the width of the car subtended about 6° of visual angle. The cars were presented so that the lower horizontal line was in the same position on every trial, centered horizontally on the screen. Participants were run individually in dimly lit, sound-dampened booths.

On each trial, the stimulus remained visible for a maximum of 30 seconds. Lack of response after 30 seconds was counted as an error. Corrective feedback appeared immediately after the participant's response. The stimulus remained visible during the feedback. The feedback appeared in three successive parts, with each part remaining visible during later parts. First, the word 'CORRECT' or 'WRONG' appeared above the stimulus, accompanied by a tone if wrong. After 500 ms the word was followed by a statement, "this takes route F(J)". Finally, after 2000 ms, the participant's percent correct over the last eight trials was displayed for 1000 ms. This extensive feedback was supplied to facilitate and motivate learning, as it was discovered in pilot experiments that participants found this to be a difficult task. After the feedback display, there was an intertrial interval of approximately 750 ms.

2.1.3. Participants. A total of 308 volunteers received partial credit toward an introductory psychology course at Indiana University. Data from participants who did not learn the initial classification were discarded because, for these participants, performance in the second phase of learning would not be based on prior learning in the first phase. Before conducting the experiment, a criterion for learning was established, which required that a participant make no more than two errors in the last 16 trials of the initial phase (trials 161–176) for their data to be included in the analysis. To guard against outlying data from learners who were atypically fast, data from participants who made three or fewer errors in the first 16 trials were also excluded. Appendix B reports results including these unusually fast learners. The effects are all the same when the fast learners are included, but they are excluded from the main analysis in order to maintain the counterbalancing of physical to abstract dimensions. Participants were assigned randomly to one of the four shift conditions, and the experiment continued until each condition had a total of 60 acceptable participants, spanning the counterbalanced assignment of physical to abstract dimensions. For the reversal-shift condition, 82 participants were run, seven were too fast, 14 were too slow, and one was discarded because he interrupted the session to ask a question. For the relevant-shift condition, 71 were run, with three too fast, seven too slow, and one was discarded because it was an accidental duplicate of a previously run condition. For the irrelevant-shift, 78 were run, with five too fast and 13 too slow. For the compound-shift, 76 were run, with four too fast, 11 too slow, and one was discarded because he took pencil-and-paper notes during the experiment, despite our efforts to put distracting stimuli such as book bags in a corner of the booth behind the participant.

2.2. Results

2.2.1. Shift learning. Accuracy as a function of training block is shown in Figure 3. Evidently, the type of shift had a large effect on ease of learning. For purposes of statistical analysis, the proportion of correct responses in the shift phase was computed for every subject individually. The mean proportion correct differed significantly between the four shift types, $F(3, 236) = 31.06$, $\text{MSW} = 0.0169$, $p < 0.0001$. Post-hoc comparisons of groups showed that reversal shift performance was better than relevant shift, $t(118) = 2.96$, SE diff $= 0.016$, $p = 0.004$

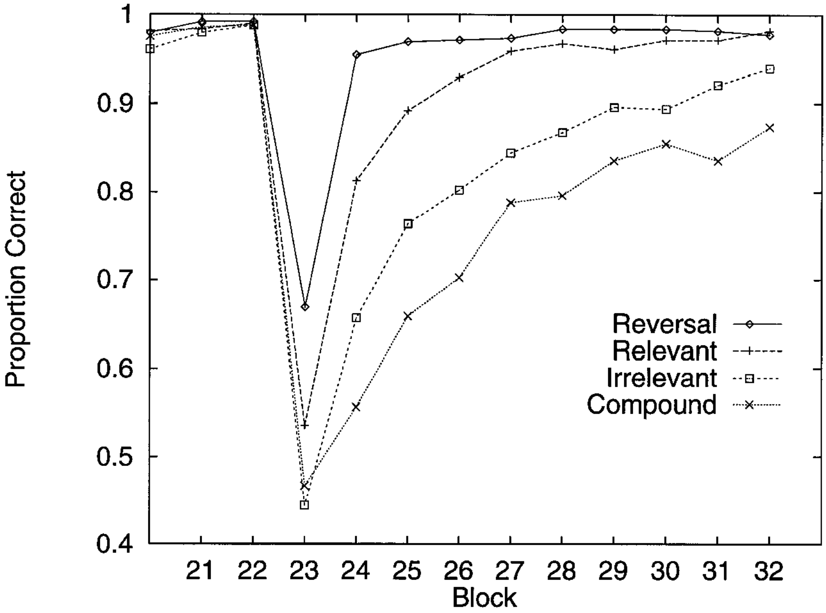

Figure 3. Results from the shift phase of the experiment with human learners. A block consisted of one presentation of each of the eight stimuli.

two-tailed; relevant shift performance was better than irrelevant shift, $t(118) = 3.65$, SE diff $= 0.026$, $p < 0.0001$ two-tailed; and irrelevant shift performance was better than compound shift, $t(118) = 2.23$, SE diff $= 0.030$, $p = 0.028$ two-tailed.

2.2.2. Initial learning. It is conceivable that the differences in learning speed between the four shift types was due to chance differences in the generic learning speed of the particular participants in the four groups. If this were the case, then learning speeds during the initial phase should also differ between the four groups. To test this possibility, the proportion of correct responses in the initial phase was computed for every subject, and an ANOVA showed that the mean proportion correct was not significantly different across the four groups, $F(3, 236) < 1$. The same conclusion is true when we restrict the analysis to the first 10 blocks of training, the same number of blocks as in shift training, $F(3, 236) = 1.136$, MSW $= 0.0209$, $p = 0.335$.

Initial learning speed was significantly slower than shift learning. Considering only the slowest of the shifts, i.e. the compound shift, for each subject the proportion correct in all 10 blocks of the shift phase was subtracted from the proportion correct in the first 10 blocks of initial learning. The mean difference across subjects was 0.0641, which was significantly greater than zero, $t(59) = 2.49$, SD $= 0.200$, $p = 0.016$ two-tailed.

2.2.3. Subproblem analysis. In any of the shifts other than reversal, half of the exemplars had their categorization changed and half of the exemplars had their categorization unchanged. It is of interest to know whether participants retained high accuracy on unchanged exemplars during the shift phase. For example, suppose that at a given point after the shift, overall performance is at 75% correct. This level of performance could be attained because accuracy of the four unchanged exemplars is perfect and accuracy on the four changed exemplars is at chance, or the overall 75% accuracy could instead be attained because performance on all eight exemplars is at 75% correct (or because of an infinite number of other combinations). Examining shift performance in terms of changed and unchanged exemplars is known as 'subproblem analysis' (Tighe *et al.*, 1971).

Figure 4 shows that performance on unchanged exemplars dropped dramatically from perfect accuracy, but not all the way to chance (50%). Throughout relearning, performance on unchanged exemplars remained better than performance on changed exemplars. The lapse on unchanged exemplars contradicts one extreme hypotheisis, viz., that participants exclusively memorized isolated exemplars.

2.3. Discussion

The relative difficulty of the four shift types cannot be accounted for by (a) differences in the generic learning speeds of the participants in the four groups, because of the four groups learned the initial classification at comparable speeds, (b) the number of dimensions that had their relevance shifted, because this mispredicts the order of the irrelevant and compound shifts, (c) the number of exemplars that had their classification changed, because the relevant, irrelevant and compound shifts all had four of eight exemplars change classification or (d) the number of dimensions relevant to the new classification, because the reversal

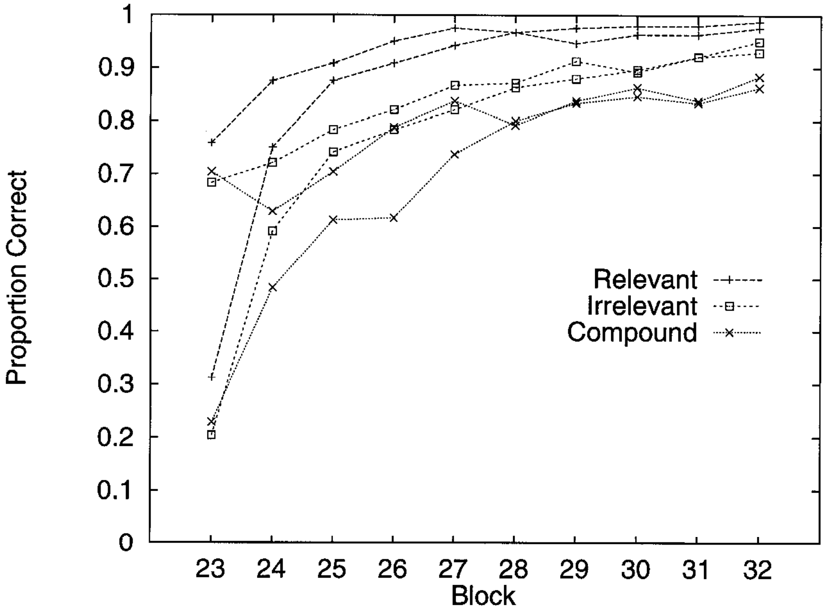

Figure 4. Subproblem analysis of data from the experiment with human learners. For each shift type, there are two curves: the upper curve is the proportion correct for unchanged exemplars; the lower curve is the proportion correct for changed exemplars.

and compound shifts both have two relevant dimensions and the relevant and irrelevant shifts both have one relevant dimension.

The relative difficulty of the four shifts also disproves theories that predict that a compound categorization will be learned more quickly than a component categorization after previous compound categorization. The data might appear to conflict with previous experiments that did find such an enhancement of compound categorization learning, but a rapprochement is available. Unlike the experiment reported in this article, the previous experiments of Zeaman and House (1974) and Barnes *et al.* (1978) used novel stimulus values, or novel variations on dimensions, in the shift phase. These manipulations might have encouraged the participants to abstract away from the specific values of the initial learning, and transfer only the most abstract structure of initial learning to the novel aspects of the shift-phase stimuli.

Despite the success of the ALCOVE model in capturing many aspects of human category learning (Choi *et al.*, 1993; Kruschke, 1992, 1993a; Nosofsky & Kruschke, 1992; Nosofsky *et al.*, 1992, 1994), it cannot account for the present results, and especially cannot capture the rapidity of learning reversal shifts. A variety of other connectionist models suffer the same failure in this situation, including the configural-cue model of Gluck and Bower (1988).

One explanation of the results is that people perseverate in their attention to dimensions and in their internal category structures. As argued by Kendler and Kendler (1962), internal category structures could facilitate rapid remapping of categories to overt responses. Stubborn attention to dimensions is needed to account for the advantage of relevant shift over irrelevant shift.

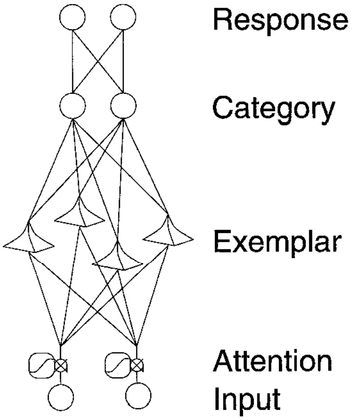

Figure 5. The AMBRY model. Only two input dimensions are illustrated. Each input node activation is modulated by a corresponding attention node, with attention determined by a sigmoidal function of contextual bias. Exemplar nodes are shown as 'pyramids' to indicate the shape of their activation function. Category nodes sum the weighted activation of the exemplar nodes. Response nodes sum the weighted activation of the category nodes.

3. AMBRY: A Connectionist Model

The principles of perseverating dimensional attention and internal category structures were implemented in a connecionist model, called AMBRY because it is a variant of the ALCOVE model (Kruschke, 1992, 1993a,b). (An ambry is a special kind of alcove.) The architecture of AMBRY is illustrated in Figure 5. One way that AMBRY extends ALCOVE is by adding another layer of nodes at the output end, reflecting associations between internal categories and overt responses. A second extension is making the attention strengths on input dimensions be sigmoidal functions of contextual bias.

3.1. Formal Description

In AMBRY, stimulus dimensions are assumed to be continuous and interval-scaled (a case of which is binary-valued), and each dimension is encoded by a separate input node such that if ψ_i is the psychological scale value of the stimulus on dimension i, then the activation of input node i is

$$a_i^{in} = \psi_i \tag{1}$$

where the superscript '*in*' indicates that this is an input node. Because all the dimensions were counter-balanced in the experimental procedure, it was assumed for simplicity that each dimension had the same salience or range of scale values with $\psi = 1.0$ for one of the binary values and $\psi = 2.0$ for the other.

There is one node established for each training exemplar. All eight exemplar nodes were included at the beginning because human participants were shown all eight stimuli as part of their initial instructions (full text of instructions is in Appendix A). The activation of an exemplar node corresponds to the psychological similarity of the current stimulus to the exemplar represented by the node. Similarity drops off exponentially with distance in psychological space, as suggested by Shepard (1987), and distance is computed using a city-block metric for psychologically separable dimensions (Garner, 1974; Shepard, 1964). Each exemplar node is significantly activated by only a relatively localized region of input space, i.e. it has a small 'receptive field'. Formally, the activation value is given by

$$a_j^{ex} = \exp\left(- c \sum_i \alpha_i \left| h_{ji} - a_i^{in} \right| \right) \qquad (2)$$

where the superscript '*ex*' indicates that this is an exemplar node, where c is a constant called the specificity that determines the overall width of the receptive field, where α_i is the attention strength on the ith dimension, and where h_{ji} is the scale value of the jth exemplar on the ith dimension. Because stimulus activation are either 1.0 or 2.0, the values of h_{ji} are either 1.0 or 2.0. Increasing the attention strength on a dimension has the effect of stretching that dimension, so that differences along the dimension have a larger influence on the similarity. This attentional flexibility is useful for stretching dimensions that are relevant for distinguishing the categories, and shrinking dimensions that are irrelevant to the category distinction (Kruschke, 1992, 1993a; Nosofsky, 1986).

In a previous article (Kruschke, 1992, p. 40), I suggested modifying the attention strengths in ALCOVE. The modification is motivated three ways. First, because negative attention strengths have no clear psychological interpretation, it would be appropriate to keep the attention strengths non-negative without just arbitrarily clipping them at zero. Second, it might be helpful to keep the attention strengths bounded above, again in some way other than arbitrary clipping, to express a possible capacity constraint on attention strength. (Nosofsky (1986) described a capacity constraint in which the total attention across all dimensions sums to a constant value, so that the dimensions are effectively competing for attention. The capacity constraint here is different, as it imposes an upper bound on individual dimensional attention strengths independent of other dimensions.) Third, and most importantly in the present context, people's apparent perseveration of attention to previously relevant dimensions might be accounted for if attention that is very strong or very weak tends to stay that way, even when the classification feedback changes. How might these desiderata be met? One possibility is to let the attention strength α_i be a function of some underlying variable β_i, rather than a primitive in the formalization:

$$\alpha_i = 1/(1 + \exp[-\beta_i]) \qquad (3)$$

This sigmoidal function in equation (3) bounds α_i in the interval (0, 1). Its slope is small when α_i is near its extreme values, and so learning by gradient descent should be slow when attention is extreme; i.e. attention strengths should be relatively reluctant to change from extreme values.

The values of β_i in equation (3) are initialized such that attention is evenly distributed across the dimensions and sums to 1.0. In the present application, there are three dimensions, so initially $\beta_i = -\ln(3-1)$, hence $\alpha_i = 1/3$. The variables β_i can be interpreted as learned contextual biases, or strengths of association between situational context and dimensional attention. When the situational context remains relatively constant, as in the shift-learning experiment, then the contextual bias to certain dimensions will persist when the categorization changes. When the situational context changes drastically, there will less transfer of attentional bias to the new context.

Activation from the exemplar nodes is propagated to category nodes, which correspond to internal representations of categories. There is one category node per category label. The activation of the *k*th category node is determined by a linear combination of exemplar-node activations:

$$a_k^{cat} = \sum_{\substack{ex \\ j}} w_{kj}^{cat} \, a_j^{ex} \tag{4}$$

where w_{kj}^{cat} is the association weight to category node k from exemplar node j. The exemplar-to-category association weights are initialized at zero.

The activation of category nodes is then propagated to an additional layer of response nodes. This formalizes the idea that internal category knowledge can remain relatively fixed while the response to a category changes. This extra response-node layer is crucial for capturing the rapidity of learning reversal shift. There is one response node per category label, with activation given by:

$$a_r^{resp} = \sum_{\substack{cat \\ k}} w_{rk}^{resp} \, a_k^{cat} \tag{5}$$

where w_{rk}^{resp} is the association weight to response node r from category node k. The category-to-response association weights are initialized such that $w_{rk}^{resp} = 1.0$ if $r = k$ and $w_{rk}^{resp} = 0.0$ otherwise.

Response node activations (exponentiated to make them non-negative) are mapped to response probabilities using the Luce (1963) choice rule (a.k.a. the softmax rule in computer science):

$$\Pr(R) = \exp(\phi a_R^{resp}) \Big/ \sum_{\substack{resp \\ r}} \exp(\phi a_r^{resp}) \tag{6}$$

where ϕ is a scaling constant. In other words, the probability of classifying the given stimulus into category R is determined by the magnitude of category R's response activation relative to the sum of all response activations.

The dimensional attention strengths and the association weights are learned by gradient descent on sum-squared error, as used in standard backpropagation (Rumelhart *et al.*, 1986). Each presentation of a training exemplar is followed by feedback indicating the correct response, just as in the categorization experiments with human participants. The feedback is coded in AMBRY as teacher values, t_r, given to each response node. For a given training exemplar and feedback, the error generated by the model is defined as

$$E = \frac{1}{2} \sum_{\substack{resp \\ r}} (t_r - a_r^{resp})^2 \tag{7}$$

where the teacher values are defined in these simulations as $t_r = +1$ if the stimulus is a member of category r, and $t_r = 0$ if the stimulus is not a member of category r.

Upon presentation of a training exemplar to AMBRY, the association weights and attention strengths are changed so that the error decreases. Following Rumelhart *et al.* (1986), they are adjusted proportionally to the (negative of the) error gradient. Evaluating the derivatives leads to the following learning rules:

$$\Delta w_{rk}^{resp} = \lambda_r (t_r - a_r^{resp}) a_k^{cat} \tag{8}$$

$$\Delta w_{kj}^{cat} = \lambda_c \left(\sum_{\substack{resp \\ r}} (t_r - a_r^{resp}) w_{rk}^{resp} \right) a_j^{ex} \tag{9}$$

$$\Delta\beta_i = -\lambda_\alpha \sum_{\substack{ex\\j}} \left[\sum_{\substack{cat\\k}} \left(\sum_{\substack{resp\\r}} (t_r - a_r^{resp}) w_{rk}^{resp} \right) w_{kj}^{cat} \right] a_j^{ex}\, c \left| h_{ji} - a_i^{in} \right| \alpha_i (1 - \alpha_i) \tag{10}$$

where the λs are constants of proportionality ($\lambda > 0$) called 'learning rates'.

The category-to-response weights are supposed to reflect a psychological bias toward one-to-one mappings of categories to responses, so the category-to-response weights are renormalized after every trial, as follows. First, any negative category-to-response weights are set to zero. Next, each weight is divided by the sum of the weights fanning into the response node. In this way the resulting sum of fan-in weights is 1.0. This renormalization is crucial to the model's ability to learn reversal shift quickly, because the renormalization is crucial for implementing the principle of rapidly permuting the mapping from categories to responses.

Learning in AMBRY proceeds as follows. For each presentation of a training exemplar, activation propagates to the response nodes and choice probabilities are recorded. Then the teacher values are presented and error is propagated down the network. The association weights and attention strengths are then adjusted. In fitting AMBRY to human learning data, there are five free parameters: the fixed specificity c in equation (2); the probability mapping constant ϕ in equation (6); the category-to-response association weight learning rate λ_r in equation (8); the exemplar-to-category association weight learning rate λ_c in equation (9); and the attention strength learning rate λ_α in equation (10).

3.2. Fit Results

The model was fit to frequencies of correct or wrong responses in the last 12 blocks of training, which included the last two blocks of initial learning plus all 10 blocks of shift learning. Therefore the model was forced to achieve high accuracy at the end of initial training, plus fit the relative rates of relearning the four shift types. Specifically, the model was fit to the frequencies in the $4 \times 12 \times 2$ table defined by crossing shift type, training block and response accuracy (correct or wrong). The model was trained on the same particular sequences as seen by the human participants, i.e. 60 sequences per shift type. Because the training sequence determined the frequency of occurrence of each shift type in each block, the degrees of freedom (df) in the data were $4 \times 12 \times (2 - 1) = 48$. The full AMBRY model, with five freely estimated parameters, therefore left 43 df. Best fitting parameter values were searched for with simulated annealing (Corana *et al.*, 1987; Goffe *et al.*, 1994) and hill-climbing on the log-likelihood measure of discrepancy, $G^2 = 2\Sigma_i f_i \ln(f_i/\hat{f}_i)$, where f_i is the observed frequency in cell i, $\hat{f}_i$ is the predicted frequency in cell i, and the sum is taken over all cells in the table (Wickens, 1989).

3.2.1. Fit to shift learning. AMBRY nicely showed the relative ease of the four shift types, as displayed in Figure 6. It rapidly learns reversal shift, learns relevant shift faster than irrelevant shift, and learns compound shift most slowly. Despite the good qualitative fit, the model can be statistically rejected. The model deviates from the data most in the early blocks of the shift period, especially block 24. It can be seen in Figure 6 that performance does not rise as quickly in AMBRY as in human learners. Nevertheless, AMBRY is the only model, of which I am aware, that can even qualitatively capture the relative difficulty of these four shifts.

Without dimensional attention learning, the model cannot show the advantage of relevant shift over irrelevant shift. Figure 7 shows the best fit of AMBRY

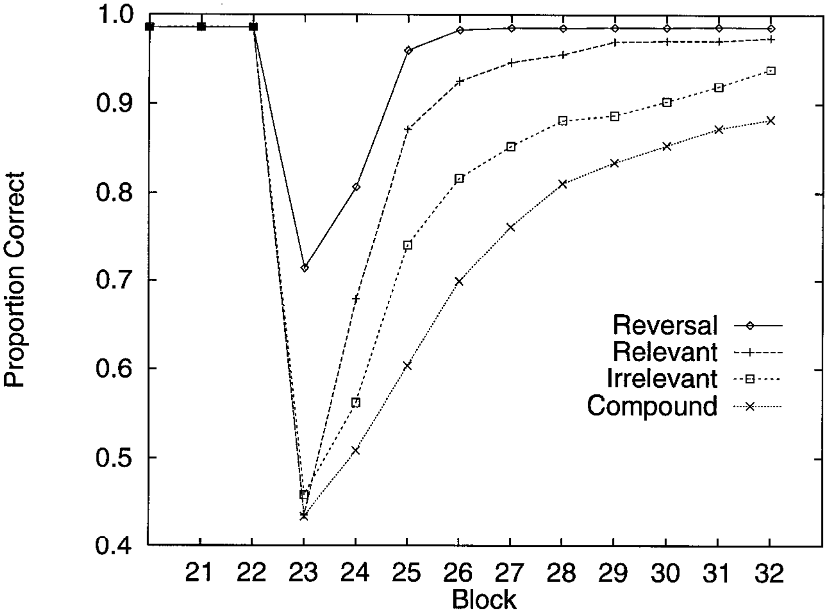

Figure 6. Best fit of AMBRY to the data shown in Figure 3. (G^2(43, $N = 23{,}040$) = 224.67, $\phi = 4.283$, $c = 4.683$, $\lambda_r = 0.8553$, $\lambda_c = 1.067$, $\lambda_\alpha = 3.637$.)

without dimensional attention learning (λ_α fixed at zero), where it can be seen that there is little separation of the performance curves for relevant and irrelevant shift. The reason for this problem is clear: without dimensional attention learning, the different dimensions needed in the relevant and irrelevant shifts are equally attended to in the initial phase, so that relevant shift has no strong advantage. The small separation between the performance curves for relevant and irrelevant shift is caused by similarities of the changed exemplars. In relevant shift, the four changed exemplars are immediate neighbors of each other (see Figure 1), and so relearning of one can benefit the others due to their similarity. In irrelevant shift, the four changed exemplars are maximally spread out at opposite corners of the 'cube' in stimulus space (see Figure 1), so learning of individual changed exemplars does not help learning of other changed exemplars.

Without category-to-response weight learning, the model cannot show the advantage of reversal shift over the other shift types. Figure 8 shows the best fit of AMBRY without category-to-response association weight learning (λ_r fixed at zero). In particular, relevant shift is learned faster than reversal shift. (The same result is also true for the original version of ALCOVE.) Dimensional attention learning benefits the relevant shift because one of the two dimensions attended to in the initial space can be collapsed. Consequently, learning of exemplar-to-category associations goes faster because of cooperation between exemplars in close proximity in the collapsed space. This does not occur in reversal shift, however, because both of the initially relevant dimensions remain relevant.

3.2.2. Predictions for initial learning. Human participants learned the shift phase faster than the initial phase, even for the compound shift. This was not true of AMBRY (nor of any other backprop-like models I have tested, such as ALCOVE).

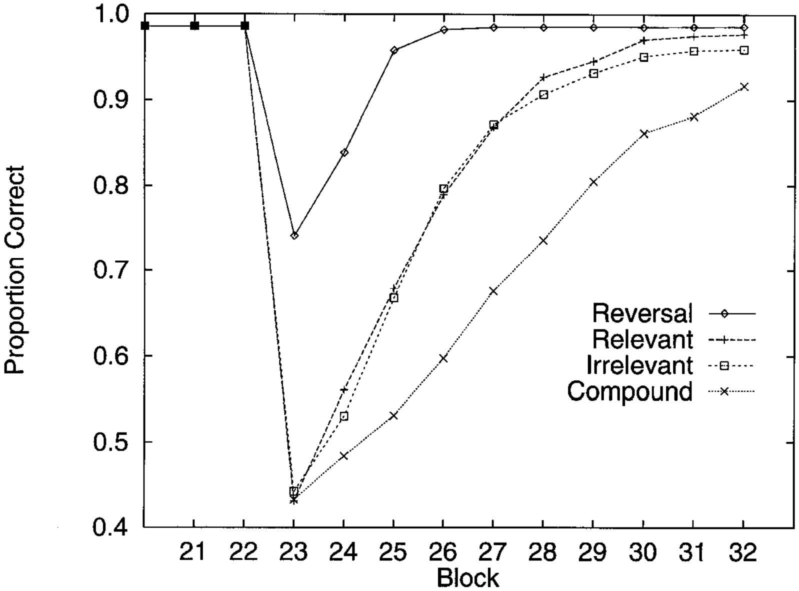

Figure 7. Best fit of AMBRY without dimensional attention learning to the data shown in Figure 3. Notice that there is little separation of the curves for the relevant and irrelevant shifts. ($G^2(44, N = 23{,}040) = 745.49$, $\phi = 4.220$, $c = 5.683$, $\lambda_r = 0.9198$, $\lambda_c = 0.9176$, $\lambda_\alpha = 0.0$ (fixed).)

With the parameter values that best fit the shift phase, AMBRY learned the initial phase very quickly, reaching asymptotic performance by the sixth block. This is understandable: in the first phase, the model does not need to overcome any prior weights or biases, whereas in the shift phase it must 'undo' prior learning.

The model could be modified to reflect better the relative learning speeds of initial and shift phases. What the proper modification should be depends on a better understanding of human learning. Loosely speaking, during the initial phase it is plausible that human learners are 'growing accustomed' to the task environment, the stimuli, the response keys and so on, and probably people are also learning to narrow their attention to the specific dimensions of variation between exemplars. This might be partially addressed in the model by setting the initial values of dimensional attention to very low levels, but might require inclusion of other mechanisms.

3.2.3. Predictions for subproblem analysis. AMBRY's predictions for performance on changed and unchanged exemplars are shown in Figure 9. Although AMBRY is an exemplar-based model, performance on the unchanged exemplars drops dramatically during the initial blocks of the shift. This is caused by the partial activation of exemplar nodes that are similar to the presented stimulus: when a stimulus with a changed categorization appears, it partially activates similar exemplar nodes, some of which have unchanged categorizations, hence the association weights from unchanged exemplars are also partially changed.

Figure 9 also shows that performance on the unchanged exemplars actually drops below performance on the changed exemplars, unlike the human data. This inversion is caused by the rapid changes in the category-to-response associations.

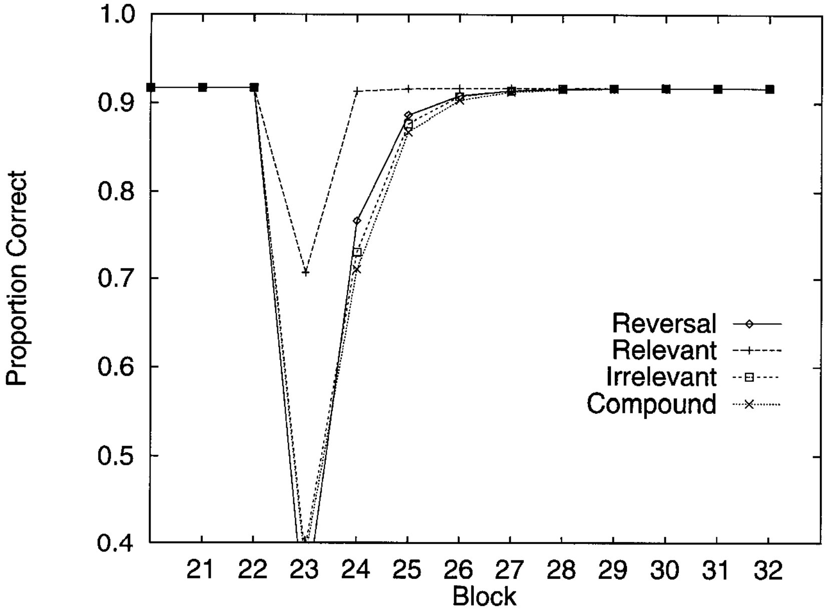

Figure 8. Best fit of AMBRY without category-to-response learning to the data shown in Figure 3. Notice that reversal shift is learned much too slowly, even slower than relevant shift. ($G^2(44, N = 23{,}040) = 2{,}038.64$, $\phi = 2.403$, $c = 7.835$, $\lambda_r = 0.0$ (fixed), $\lambda_c = 0.5954$, $\lambda_\alpha = 10.0$ (maximal allowed value).)

Indeed, when the category-to-response association learning rate is fixed at zero, no such inversion occurs (but the unchanged exemplars continue to show a dramatic drop in accuracy after the shift, so that the drop is not solely attributable to changes in the category-to-response association weights). AMBRY, apparently unlike human learners, is willing to change its category-to-response mapping at any time there is error. People apparently have more focused strategies regarding when to apply this change.

3.3. Discussion

AMBRY is able to show the correct ordering of the four shift types, unlike some previous connectionist models. It does not, unfortunately, reflect human behavior with great accuracy. One problem with AMBRY in its present formulation is that on the first reversal shift trial, not only do the category-to-response association weights change rapidly, but so do the exemplar-to-category association weights. In particular, the association weight from the presented exemplar changes its sign, and it is not rectified until the exemplar is presented again on the next block. Humans do not exhibit this effect; that is, they do not show, in the second post-shift block, non-reversed classification of the first exemplar seen in reversal shift. One modification to AMBRY that eliminates this problem is to let the activation of the category nodes be a sigmoidal function of their net input. After extensive learning in the initial phase, category node activations are near asymptote, and so the error propagated through the category nodes is small (because it is multiplied by the derivative of the activation function, $a_k^{cat}[1 - a_k^{cat}]$). While this

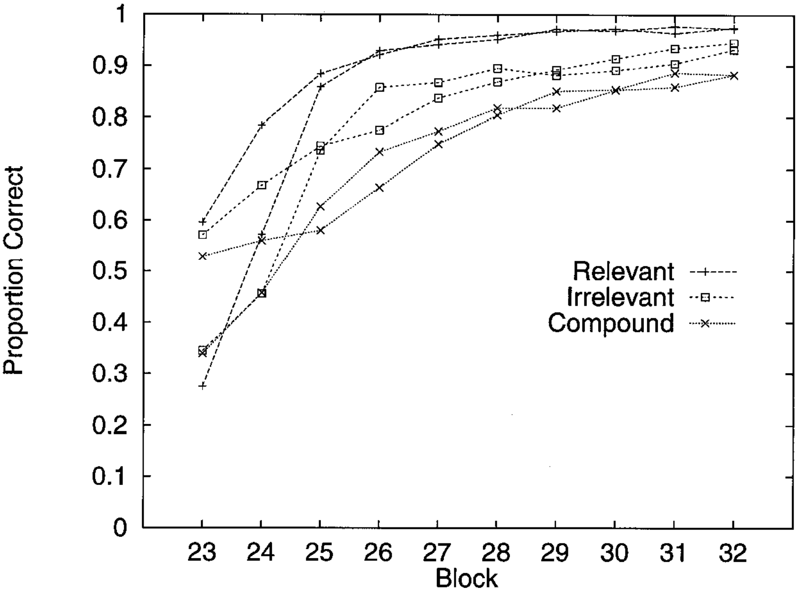

Figure 9. Subproblem analysis of predictions from AMBRY. The corresponding human data in Figure 4 were not fit directly; these predictions are from parameter values that best fit the overall proportion correct, reported in Figure 6. For each shift type, there are two curves: the curve that begins higher is the proportion correct for unchanged exemplars; the curve that begins lower is the proportion correct for changed exemplars.

modification solves the problem of over-eager changes in exemplar-to-category association weights, it introduces another problem: the error signal is too small for learning the other shift types quickly enough.

Another inadequacy of AMBRY in its present formulation appeared in the subproblem analysis, where it was seen that performance on the changed exemplars at times exceeded performance on the unchanged exemplars. What is needed is a reversal mechanism that is elicited at strategic times, rather than applied full-force on all trials.

As mentioned before, AMBRY is unable to learn the compound shift more quickly than the initial (compound) categorization. Moreover, the model has no mechanism by which it could progressively improve its performance on successively repeated shifts of the same type, as humans (and other animals) do. This might be partially addressed by adding mechanisms that improve discrimination and learning rates, but more likely it would require some sort of meta-learning device.

4. Conclusions

The experimental results are a useful benchmark for models of learning and relearning. The data constrain theories by addressing reversal shift, compound shift, and shifts to previously relevant or irrelevant dimensions in a unified setting. The AMBRY model qualitatively captured the data as a consequence of formalizing the two principles of perseverative dimensional attention learning and category-to-response remapping. When the mediating categories were omitted, the model could not show reversal shift advantage. When dimensional attention learning was omitted, the model could not show an advantage of shifting to a previously relevant dimension. The modeling presented here suggests, like much empirical work before it, that reversal shift and extradimensional shifts are learned by humans via qualitatively different mechanisms, and future models will need to reflect this difference.

Acknowledgements

This research was supported by an Indiana University Summer Faculty Fellowship and by NIMH FIRST Award 1-R29-MH51572-01. Results from the experiment were reported at the 33rd Annual Convention of the Psychonomic Society, St. Louis, 1992. William Maki and Robert Goldstone brought the issue of reversal-shift learning to the author's attention. Appreciation is extended to Lawrence Barsalou, Nathaniel Blair, James Chumbley, Michael Erickson, William Estes, Michael Fragassi, Robert Goldstone, Mark Johansen and Michael McCloskey for commenting on previous versions of this article. Michael Erickson provided valuable assistance with computer programming. Thanks to Matt Darbro, Michael Erickson, Chuck Hagen, Lisa Kercher and Jenni McDaniel for help administering the experiment.

References

Anderson, D.R., Kemler, D.G. & Shepp, B.E. (1973) Selective attention and dimensional learning: A logical analysis of two-stage attention theories. *Bulletin of the Psychonomic Society*, **2**, 272–275.

Barnes, T.R., Cassidy, E.M., Ninfa, J.M., Yago, M.M. & Barnes, M.J. (1978) Transfer of compound and component solution modes. *Memory & Cognition*, **6**, 607–611.

Choi, S., McDaniel, M.A. & Busemeyer, J.R. (1993) Incorporating prior biases in network models of conceptual rule learning. *Memory & Cognition*, **21**, 413–423.

Corana, A., Marchesi, M., Martini, C. & Ridella, S. (1987) Minimizing multimodal functions of continuous variables with the 'simulated annealing' algorithm. *ACM Transactions on Mathematical Software*, **13**, 262–280.

Garner, W.R. (1974) *The Processing of Information and Structure.* Hillsdale, NJ: Erlbaum.

Gluck, M.A. & Bower, G.H. (1988) Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, **27**, 166–195.

Gluck, M.A. & Myers, C.E. (1992) Hippocampal-system function in stimulus representation and generalization: A computational theory. *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, 390–395.

Gluck, M.A., Glauthier, P.T. & Sutton, R.S. (1992) Adaptation of cue-specific learning rates in network models of human category learning. *Proceedings of the 14th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum, 540–545.

Goffe, W.L., Ferrier, G.D. & Rogers, J. (1994) Global optimization of statistical functions with simulated annealing. *Journal of Econometrics*, **60**, 65–99.

Kendler, H.H. & Kendler, T.S. (1962) Vertical and horizontal processes in problem solving. *Psychological Review*, **69**, 1–16.

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

Kruschke, J.K. (1993a) Human category learning: Implications for backpropagation models. *Connection Science*, **5**, 3–36.

Kruschke, J.K. (1993b) Three principles for models of category learning. In G.V. Nakamura, R. Taraban & D.L. Medin (Eds), *Categorization by Humans and Machines: The Psychology of Learning and Motivation*, Vol. 29. San Diego, CA: Academic Press, pp. 57–90.

Levine, M. (1975) *A Cognitive Theory of Learning: Research on Hypothesis Testing.* Hillsdale, NJ: Erlbaum.

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

Mackintosh, N.J. (1965) Selective attention in animal discrimination learning. *Psychological Bulletin*, **64,** 124–150.

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

Nosofsky, R.M. & Kruschke, J.K. (1992) Investigations of an exemplar-based connectionist model of category learning. In D.L. Medin (Ed.), *The Psychology of Learning and Motivation*, Vol. 28. San Diego, CA: Academic Press, pp. 207–250.

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

Nosofsky, R.M., Gluck, M.A., Palmeri, T.J., McKinley, S.C. & Glauthier, P. (1994) Comparing models of rule-based classification learning: A replication of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, **22,** 352–369.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986) Learning internal representations by error propagation. In J.L. McClelland & D.E. Rumelhart (Eds), *Parallel Distributed Processing*, Vol. 1. Cambridge, MA: MIT Press, Chapter 8, pp. 318–362.

Shepard, R.N. (1964) Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, **1,** 54–87.

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

Tighe, T.J., Glick, J. & Cole, M. (1971) Subproblem analysis of discrimination shift learning. *Psychomonic Science*, **24,** 159–160.

Wickens, T.D. (1989) *Multiway Contingency Tables Analysis for the Social Sciences.* Hillsdale, NJ: Erlbaum.

Wolff, J.L. (1967) Concept-shift and discrimination-reversal learning in humans. *Psychological Bulletin*, **68,** 369–408.

Zeaman, D. & House, B.J. (1974) Interpretations of development trends in discriminative transfer. In A.D. Pick (Ed.), *Minnesota Symposia on Child Psychology*, Vol. 8. Minneapolis, MN: University of Minnesota Press, pp. 144–186.

Appendix A: Full Text of Instructions to Participants

The following instructions were displayed on the computer screen for the participant to read. They were simultaneously read aloud by the experimenter.

> This experiment investigates how people learn to distinguish and classify objects. You'll be shown simple line drawings of box cars from freight trains. Each freight car takes one of two routes, labeled F or J. Your job is to learn which route each car takes. Each time a freight car is shown, you should press the key corresponding to that car's route. After you make your response, the computer will display whether you're correct or wrong, and it will also tell you the car's actual route. The first several times you see each car you'll be guessing, but after a while you'll learn which car goes on which route.
>
> You will see examples of all eight freight cars. The cars vary in just three ways: their height, whether or not their wheels are dirty (blank) or clean (filled white), and the position of a 'door' on the car. The route a car takes is completely determined by which combination of those three features it has.
>
> Press the space bar to see the examples.

The eight exemplars were then presented in a different random order for each participant. Each exemplar was displayed for 1000 ms with a blank inter-exemplar interval of approximately 750 ms. The participant made no response to any exemplar. During this time the experimenter pointed out the three dimensions of variation on the display.

Remember that the eight cars vary in just three ways, and the route taken is completely determined by those features. During the experiment, you will respond to each car by pressing a route label, F or J, and then the computer will display the correct answer. Use the index fingers of both hands to press the keys.

You should try to learn the routes taken by each car as accurately as possible. Your percent correct for the previous 8 trials will be shown at the end of each trial (beginning on the 8th trial). At first you will just be guessing and achieve only about 50% correct, but if you try, you can work up to 100% correct.

There might be some times during the experiment when the routes taken by the freight cars change, but that will happen only very rarely. Previous experiments have shown that nearly all subjects get 100% correct by the time any routes change.

There is no emphasis on response speed; you may take up to 30 seconds to respond to each trial (but longer than that is counted as an error). You'll have one brief break during the experiment (the computer will indicate when).

Press the space bar to begin.

Appendix B: Results Including Unusually Fast Learners

The results reported in the main text excluded data from participants who did not adequately learn the initial categorization, because data from these participants would not be indicative of relearning after prior learning. The results in the main text also excluded data from some participants who learned unusually quickly. This Appendix reports results with the unusually fast learners included. For reversal shift, there were seven fast learners, so the enlarged set of data includes 67 participants. For relevant, irrelevant and compound shifts, there were 63, 65 and 64 participants in the enlarged sets, respectively. By including these additional participants, the assignment of physical to abstract dimensions is no longer fully

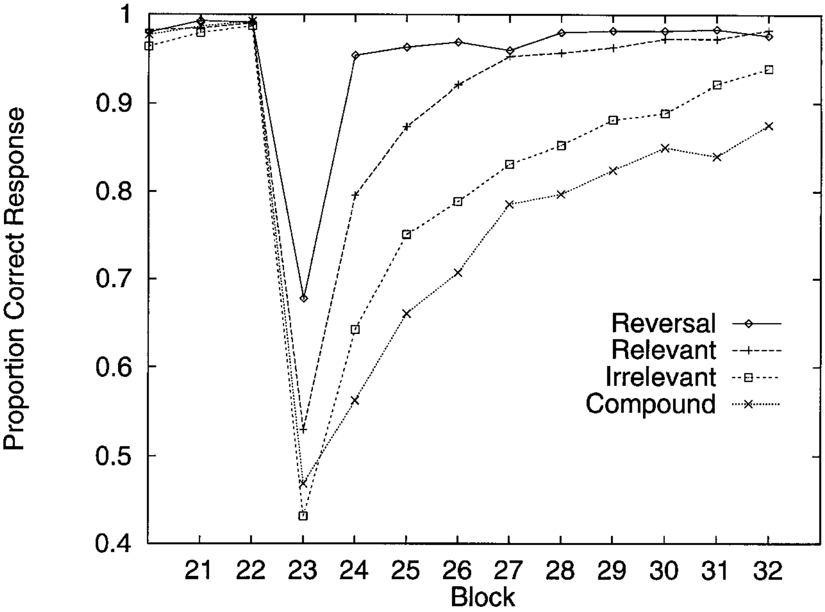

Figure 10. Results from the shift phase of the experiment with human learners, including the unusually fast learners. A block consisted of one presentation of each of the eight stimuli. Compare with Figure 3.

counterbalanced. Figure 10 shows results of the shift learning phase with the unusually fast learners included. The results are extremely similar to the results with the fast learners excluded; compare with Figure 3. The same ordering of difficulty of the four shift types is found.