

## CHAPTER FOUR

# Learning involves attention

**John K. Kruschke**

*Indiana University, Bloomington, IN, USA*

Kruschke, J. K. (2005). Learning involves attention.  
In: G. Houghton (Ed.), *Connectionist Models in Cognitive Psychology*,  
Ch. 4, pp. 113-140. Hove, East Sussex, UK: Psychology Press.

## INTRODUCTION

One of the primary factors in the resurgence of connectionist modeling is these models' ability to learn input-output mappings. Simply by presenting the models with examples of inputs and the corresponding outputs, the models can learn to reproduce the examples and to generalize in interesting ways. After the limitations of perceptron learning (Minsky & Papert, 1969; Rosenblatt, 1958) were overcome, most notably by the back-propagation algorithm (Rumelhart, Hinton, & Williams, 1986) but also by other ingenious learning methods (e.g. Ackley, Hinton, & Sejnowski, 1985; Hopfield, 1982), connectionist learning models exploded into popularity. Connectionist models provide a rich language in which to express theories of associative learning. Architectures and learning rules abound, all waiting to be explored and tested for their ability to account for learning by humans or other animals.

A thesis of this chapter is that connectionist learning models must incorporate rapidly shifting selective attention and the ability to learn attentional redistributions. This kind of attentional shifting is not only necessary to mimic learning by humans and other animals, it is also a highly effective and rational solution to the demands of learning many new associations as quickly as possible. This chapter describes three experiments (one previously published and two new) that demonstrate the action of attentional learning. All the results are fitted by connectionist models that shift and learn

attention, but the results cannot be fitted when the attention mechanisms are shut off.

### Shifts of attention facilitate learning

A basic fact of learning is that people quickly learn new associations without rapidly forgetting old associations. Presumably this ability is highly adaptive for any creature that confronts a rich and complex environment. Consider a hypothetical situation in which an animal learns that mushrooms with a round top and smooth texture are tasty and nutritious. After successfully using this knowledge for some time, the animal encounters a new mushroom with a smooth texture but a flat top. This mushroom turns out to induce nausea. How is the animal to quickly learn about this new kind of mushroom, without destroying still-useful knowledge about the old kind of mushroom? If the animal learns to associate both features of the new mushroom with nausea, then it will inappropriately destroy part of its previous knowledge about healthy mushrooms, i.e. the previous association from smooth texture to edibility will be destroyed. On the other hand, if the old association is retained, it generates a conflicting response, i.e. eating the mushroom.

To facilitate learning about the new case, it would be advantageous to selectively attend to the distinctive feature, viz. flat top, and learn to associate this feature with nausea. By selectively attending to the distinctive feature, previous knowledge is preserved, and new learning is facilitated. Not only should attention be shifted in this way to facilitate learning, but the shifted attentional distribution should itself be learned: Whenever the animal encounters a mushroom with smooth texture and flat top, it should shift attention to the flat top, away from the smooth texture. This will allow the animal to properly anticipate nausea, and to avoid the mushroom.<sup>1</sup> The third example in this chapter describes a situation in which people use exactly this kind of attentional shifting during learning. The challenge to the theorist is expressing these intuitions about attention in a fully specified model.

### Shifts of attention can be assessed by subsequent learning

The term “attention”, as used here, refers to both the influence of a feature on an immediate response and the influence of a feature on learning. If a feature is being strongly attended to, then that feature should have a strong influence on the immediate response and on the imminent learning. This latter influence of attention on learning is sometimes referred to as the feature’s *associability*. In this chapter, these two influences of attention are treated synonymously. This treatment is a natural consequence of the connectionist

models described below, but the treatment might ultimately turn out to be inappropriate in the face of future data.

Because redistribution of attention is a learned response to stimuli, the degree of attentional learning can be assayed by examining *subsequent* learning ability. If a person has learned that a particular feature is highly indicative of an appropriate response, then, presumably, the person has also learned to attend to that feature. If subsequent training makes a different feature relevant to new responses, then learning about this new correspondence should be relatively slow, because the person will have to unlearn the attention given to the now-irrelevant feature. In general, learned attention to features or dimensions can be inferred from the ease with which subsequent associations are learned. This technique is used in all three examples presented below.

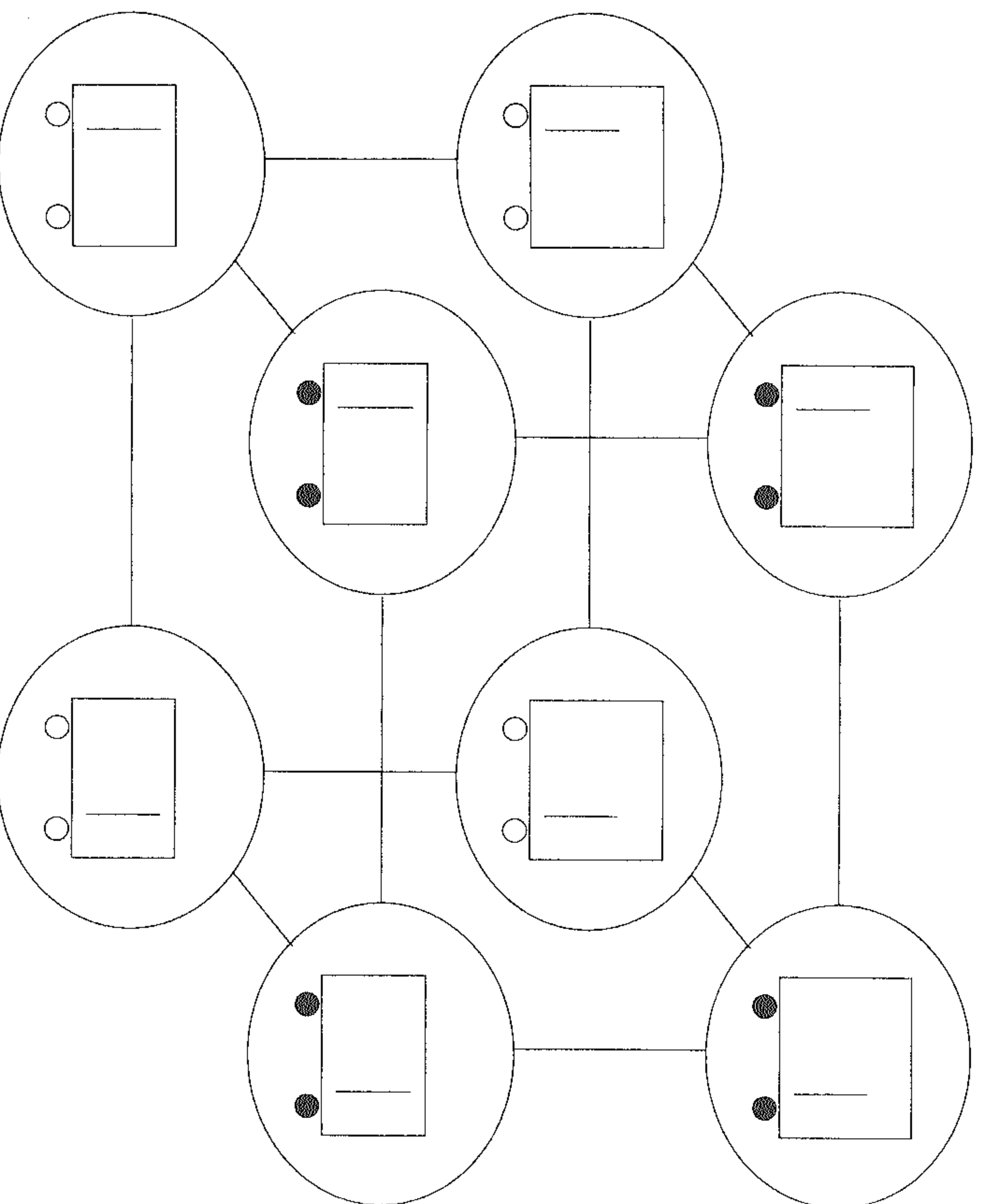
### INTRA- AND EXTRADIMENSIONAL SHIFTS

A traditional learning paradigm in psychology investigates perseveration of learned attention across phases of training. In the first phase, participants learn that one stimulus dimension is relevant to the outcome while other dimensions are irrelevant. In the second phase, the mapping of stimuli to outcomes changes so that either a different dimension is relevant or the same dimension remains relevant. The former change of relevance is called *extradimensional* shift, and the latter change is called *intradimensional* shift. Many studies in many species have shown that intradimensional shift is easier than extradimensional shift, a fact that can be explained by the hypothesis that subjects learn to attend to the relevant dimension, and this attentional shift perseverates into the second phase (e.g. Mackintosh, 1965; Wolf, 1967). In this section of the chapter, a recent experiment demonstrating this difference is summarized, and a connectionist model that incorporates attentional learning is shown to fit the data, whereas the model cannot fit the data if its attentional learning mechanism is “turned off”.

### Experiment design and results

Consider the simple line drawings of freight train box cars shown in Figure 4.1. They vary on three binary dimensions: height, door position, and wheel color. In an experiment conducted in my lab (Kruschke, 1996b), people learned to classify these cars into one of two routes. On each trial in a series, a car would appear on a computer screen, the learner would make his/her choice of the route of the car by pressing a corresponding key, and then the correct route would be displayed. During the first few trials, the learner could only guess, but after many trials, she/he could learn the correct answers.

Figure 4.2 indicates the mapping of cars to routes. The cubes in Figure 4.2 correspond with the cube shown in Figure 4.1. Each corner is marked with a

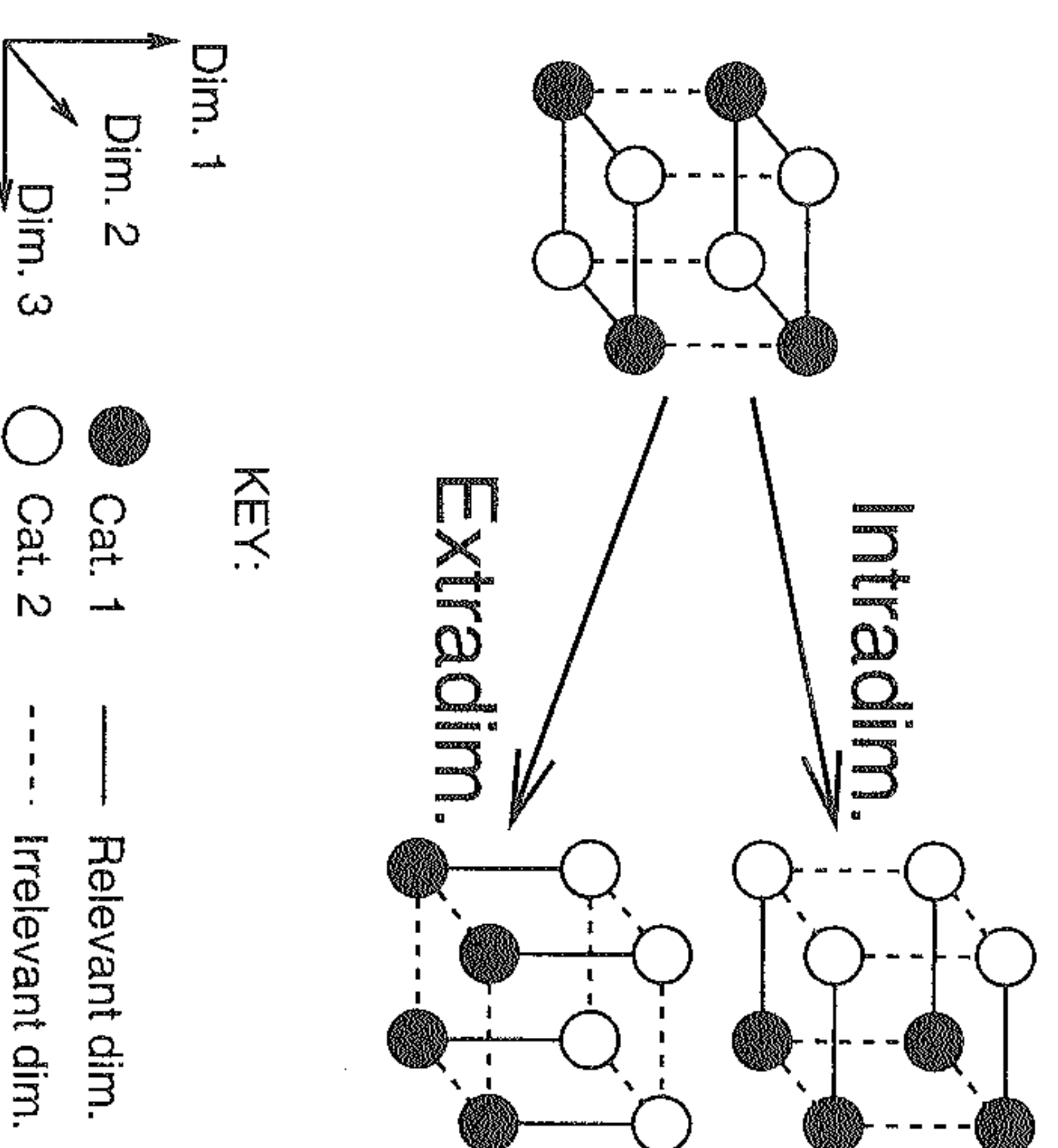


**Figure 4.1.** Stimuli used for relevance shift experiment of Kruschke (1996b). The ovals merely demarcate the different stimuli and are not part of the stimuli *per se*. The lines connecting the ovals indicate the dimensions of variation between stimuli.

disk whose color indicates the route taken by the corresponding train; in other words, the color of the disk indicates the category of the stimulus.

The left side of Figure 4.2 shows the categorization learned in the first phase of training, and the right side shows the categorization learned subsequently. In the first phase, it can be seen that the vertical dimension is irrelevant. This means that variation on the vertical dimension produces no variation in categorization: The vertical dimension can be ignored with no loss in categorization accuracy. The other two dimensions, however, are relevant in the first phase. Some readers might recognize this as the exclusive-or (XOR) structure on the two relevant dimensions.

In the subsequent phase, some learners experienced a change to the top-right structure of Figure 4.2, and other learners experienced a change to the bottom-right structure. In both of these second-phase structures only one dimension is relevant, but in the top shift this relevant dimension was one of the initially relevant dimensions, so the shift of relevance is called intradimensional, whereas in the bottom shift the newly relevant dimension was initially irrelevant, so the shift of relevance is called extradimensional. Notice



**Figure 4.2.** The structure of two types of relevance shifts. The cube at left indicates the initially learned categorization; the cubes at right indicate the alternative subsequently learned categorizations. Adapted from Kruschke, 1996b.

that the two second-phase category structures are isomorphic, so any differences in ease of learning the second phase cannot be attributed to differences in structural complexity.

This design is an advance over all previous studies of shift learning because no novel stimulus values are used in either shift. Thus, intradimensional and extradimensional shifts can be directly compared without confounded changes in novelty. In traditional studies of intradimensional shift, the shift is accompanied by introduction of novel values on the relevant dimension. For example, the initial phase might have color relevant, with green indicating category X and red indicating category Y. The only way to implement an intradimensional shift, without merely reversing the assignment of categories to colors, is to add novel colors; e.g. yellow indicates X and blue indicates Y. Unfortunately, if novel features are added to the initially relevant dimension, it might be the case that differences in learnability of the dimensions were caused by differences in novelty. If novel features are added to both dimensions, it might be the case that differences in learnability are attributable to differences in degree of novelty, or differences in similarity of the novel values to the previous values, and so forth (Slamecka, 1968). This new design solves these problems by making the initial problem involve two relevant dimensions, and no novel values at all in the shift phase.<sup>2</sup>

Human learning performance in this experiment is shown in Figure 4.3. It can be seen that people learned the intradimensional shift much faster than the extradimensional shift [ $t(118) = 3.65$ ,  $SE_{diff} = .026$ ,  $p < .0001$  two-tailed].

Notice that the advantage of the intradimensional shift over the

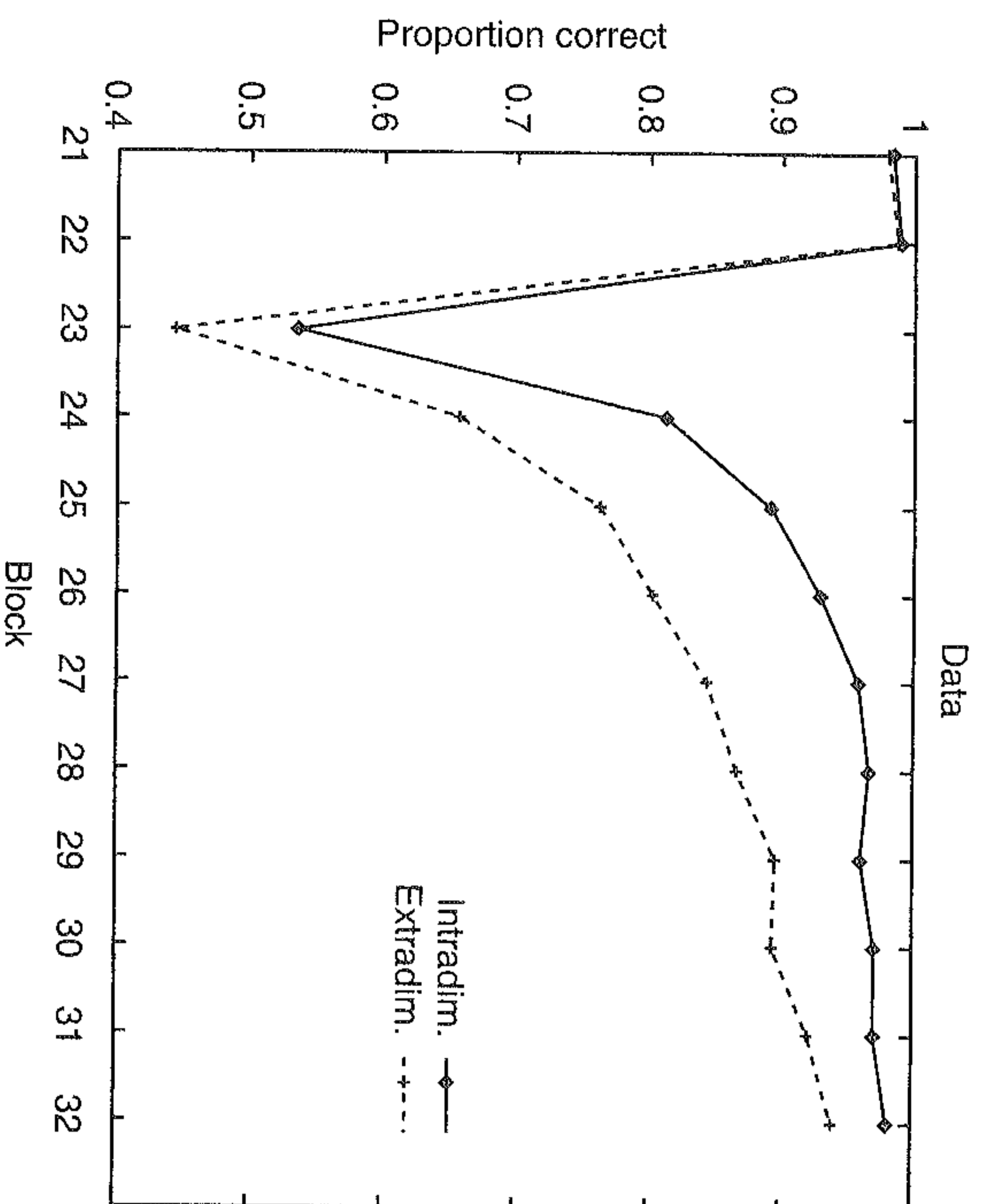


Figure 4.3. Results of the relevance shift experiment. Adapted from Kruschke, 1996b.

extradimensional shift cannot be explained by the number of exemplars that changed their route, because in both shift types there were four exemplars that changed their route. Another possible explanation for the difference is that only one dimension changed its relevance in the intradimensional shift, but all three dimensions changed their relevance in the extradimensional shift. This explanation is contradicted by results from another condition in the experiment (not summarized here), in which only two dimensions changed their relevance but the learning was even more difficult than the extradimensional shift.

This advantage of intradimensional over extradimensional shift has been found in many previous studies in many other species, but the results here are particularly compelling because the design involved no confounded variation of novelty. This robust difference should be addressed by any model of learning that purports to reflect learning by natural intelligent organisms.

### A connectionist model with attentional learning

The advantage of intradimensional shift over extradimensional shift suggests that there is learned attention to dimensions. A model of learning should implement this explanatory principle. Any model of the relevance-shift experiment will also need to be able to learn the XOR category structure in the first phase of training. This structure is nonlinear in the two relevant

dimensions, meaning that no simple additive combination of the two relevant dimensions can accurately compute the correct categories. Instead, conjunctive combinations of dimensional values must be encoded in the model. There has been much research that suggests that people can and do encode configurations of values, also called *exemplars*, during learning (e.g. Nosofsky, 1992). The model to be fitted to the shift-learning data formalizes this notion of exemplar representation, along with the notion of learned attention to dimensions.

The model fit to these data was called AMBRY by Kruschke (1996b) because it is a variant of the ALCOVE model (Kruschke, 1992). The architecture of (part of) AMBRY is shown in Figure 4.4. All aspects of the model have specific psychological motivations, and formalize explicit explanatory principles. Because of this correspondence between model parts and explanatory principles, the principles can be tested for their importance by excising the corresponding aspect of the model. In particular, the attentional mechanism can be functionally removed, and the restricted model can be tested for its ability to fit to data.

#### Activation propagation

In AMBRY, each dimension is encoded by a separate input node. If  $\psi_i$  denotes the psychological scale value of the stimulus on dimension  $i$ , then the activation of input node  $i$  is simply that scale value:

$$a_i^{\text{in}} = \psi_i \quad (1)$$

Because the experiment counter-balanced the assignment of physical dimensions in Figure 4.1 to abstract dimensions in Figure 4.2, the dimensional values were simply assumed to be 1.0 and 2.0; e.g. for the short car,  $\psi_{\text{height}} = 1.0$ , and for the tall car,  $\psi_{\text{height}} = 2.0$ .

There is one exemplar node established for each of the eight cars. 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 argued by Shepard (1987), and distance is computed using a city-block metric for psychologically separable dimensions (Garner, 1974; Shepard, 1964). An exemplar node is significantly activated only by stimuli that are fairly similar to the exemplar represented by the node. In other words, each exemplar node has a limited “receptive field” in stimulus space. Formally, the activation value is given by:

$$a_j^{\text{ex}} = \exp(-c \sum_i \alpha_i |\psi_{ji} - a_i^{\text{in}}|) \quad (2)$$

