COGNITIVE PSYCHOLOGY **32,** 1–48 (1997)
ARTICLE NO. CG970644

The Abstraction of Intervening Concepts from Experience with Multiple Input–Multiple Output Causal Environments

JEROME BUSEMEYER

Purdue University

MARK A. MCDANIEL

University of New Mexico

AND

EUNHEE BYUN

Purdue University

The purpose of this article is threefold: (a) introduce a new paradigm for investigating how intervening concepts are learned, (b) report four new experiments that provide converging evidence for the acquisition of intervening concepts, and (c) propose a simple associative learning mechanism to account for the results. The new paradigm utilizes a stimulus–response–feedback task in which subjects learn trial by trial how a multivariate set of inputs maps into a multivariate set of outputs. The first two experiments use evidence based on a principal component analysis to replicate the finding that intervening-concept learning occurs spontaneously, but only in environments that contain an intervening factor. The next experiment provides a second converging line of evidence for this conclusion by showing that subjects can use an intervening concept to make accurate inferences to a new fourth output during a transfer test. The last experiment provides a third line of evidence by showing that subjects can use an intervening concept to make accurate inferences from a new fourth input. The results are explained by a hidden-unit connectionist learning mechanism that includes both accuracy and parsimony as learning objectives.

One of the hallmarks of human intelligence is the ability to form abstract concepts from experience with specific instances. Questions concerning the nature of human conceptual learning have maintained the attention of experimen-

This work was funded by NIMH Grant MH47126. Each author made an equal contribution to this article. We thank Steve Walters for programming the stimuli used in Experiment 2. Correspondence regarding this article should be sent to Jerome R. Busemeyer, Department of Psychological Sciences, Purdue University, 1364 Psychology Building, West Lafayette, IN 47907-1364. Email: jbuse@psych.purdue.edu.

0010-0285/97 $25.00

tal psychologists for over 70 years (Hull, 1920; Bruner, Goodnow, & Austin, 1956; Hunt, 1962; Bourne, 1966; Levine, 1975; Rosch & Lloyd, 1978; Smith & Medin, 1981; Neisser, 1987). This period has seen a progression in the types of experimental paradigms used to investigate concept learning and a concomitant accruement in our basic understanding of the very nature of human concepts.

Experimental research in this area started out by defining concepts as *well-defined categories* (e.g., Bruner et al., 1956; Hunt, 1962; Bourne, 1966; Levine, 1975). The exemplars of a category were constructed from a small number of perceptual attributes (e.g., form, color, size) that varied according to a small number of features (e.g., red, blue, green, for the color attribute). A positive instance of a category was defined according to a simple logical rule, such as a conjunction or disjunction of features (e.g., red and square). A shift in research occurred in the 1970s by extending the definition of concepts to *ill-defined* or *fuzzy categories* (Rosch & Lloyd, 1978; Smith & Medin, 1981). Many natural categories such as ''games'' or ''furniture'' cannot be defined by a simple rule based on a small set of features, and the members of these categories seem to have gradients of typicality, rather than a sharp all-or-none category boundary for membership. Researchers became interested in constructing exemplars from more complex multidimensional stimuli (e.g., a family of faces varying continuously according to nose length, smile curvature, and distance between eyes). Currently, another shift in research is underfoot in which concepts are viewed as categories of objects organized together by theoretical or causal principles (Murphy & Medin, 1985; Neisser, 1987). For example, the categorization of animals and plants by modern biologists is based on genetic principles rather than the physical similarities of the category members.

Despite the paradigm changes and changes in theoretical emphasis during these 70 years of research, experimental investigations of concept learning have been limited to *categorization* (deliberately, e.g., Smith & Medin, 1981, p. 8). Essentially, the type of conceptual learning studied experimentally in the laboratory reflects learning how to assign a name to a collection of things (Bruner et al., 1956, p. 1; Hunt, Martin, & Stone, 1966, p. 10). Moreover, learning in these tasks is oriented to *concept attainment* (learning an existing category) as opposed to *concept formation* (forming a new category), because the categories are defined a priori by the experimenter and are not created by the learner. Finally, the categories used tend to be *concrete,* i.e., things that can be directly experienced in the real world (e.g., a face). There has been good reason for imposing the limitations just mentioned. Categorization is an unquestionably important type of cognitive process, and, further, the relative simplicity of the concrete concepts studied makes them attractive for laboratory experiments. By adopting such limitations, however, the literature reveals little about more complex conceptual behavior, including acquisition of the relational concepts that may form the basis for the theoretical or causal principles that underlie categorization of concrete concepts.

The purpose of this article is (a) to introduce a new concept-learning

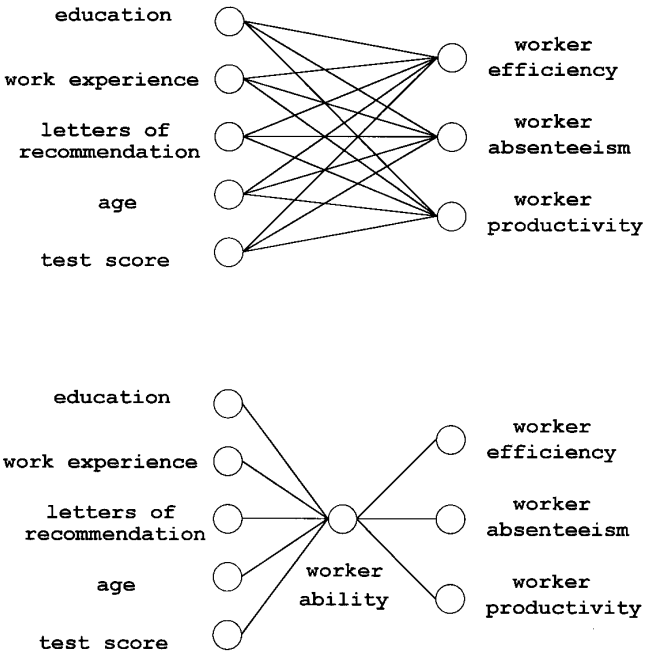


FIG. 1. Diagrams of a concept with or without an intervening variable. The top panel represents a concept without an intervening variable and the bottom panel represents a concept with an intervening variable.

paradigm that focuses on a somewhat more complex and abstract type of concept called an *intervening variable*, (b) to describe a set of initial experiments exploring the acquisition and use of intervening concepts, and (c) to briefly introduce a new type of adaptive network learning model—one constrained by considerations of parsimony—to account for intervening concept learning. Intervening concepts differ from categories in several important ways. First, intervening concepts are purely relational. Their purpose is to relate one set of variables to another, and, consequently, they play a major role in the construction of causal systems or intuitive theories about the world. Second, the learning task involves the formation or discovery of new concepts rather than the attainment of predefined concepts. Third, intervening concepts are abstract: They are not directly observable in the environment, but rather they must be created in the mind of the observer.

Figure 1 provides an example of an intervening concept ''worker ability'' borrowed from personnel psychology [there are numerous other examples (e.g., Bentler, 1980, Fig. 3; Miller, 1959, Fig. 13), but this one is presented because its structure maps onto the present experiments]. The top panel shows the situation in which the investigator attempts to explain 15 relationships between

five assessment factors and three criterial measures. The bottom panel illustrates how an intervening concept (worker ability) was created to explain the 15 input–output relationships. The basic idea is that all five assessment factors combine to form the intervening variable, and this single ''hidden'' variable then influences all of the dependent (criterial) variables (thereby reducing the number of relationships to be learned to 8). Miller used an example such as this to argue for the scientific advantage of postulating an intervening concept. By using an intervening variable, the total number of relationships that must be learned is an additive rather than a multiplicative function of the number of inputs and outputs. This produces considerable savings when both the number of inputs and the number of outputs are relatively large.

As the reader probably recognizes, the preceding example is by no means an isolated one; many others readily come to mind. In psychological testing, general intelligence (g) is a construct that intervenes between a set of test questions (inputs) and test responses (outputs). In the personality area, social stimuli (inputs) affect an intervening construct, self-esteem, which in turn influences social behavior (outputs). In experimental psychology the variables hours of deprivation, feeding on dry food, and saline injection are used to infer the intervening variable, thirst, and this intervening variable is then used to predict rate of bar pressing, volume of water drunk, and amount of quinine required to stop drinking (Miller, 1959, Fig. 13). Finally, note that all of these examples seem to satisfy the criteria for intuitive theories outlined in Murphy and Medin (1985). In each of these examples, the formation of an intervening concept serves to provide a simpler understanding of the causal relations mediating multiple inputs and outputs.

These examples document the fact that learning intervening, causal relations between multiple inputs and outputs, represents a central concept-formation process for at least one highly specialized group of people—scientists. It is also patent that intervening variables are naturally employed by people in general to understand the complex causal relations they encounter in their daily lives. The frequent occurrence of terms such as hunger, thirst, and intelligence in everyday language indicates that intervening concepts are commonly used. What is at issue is whether intervening concepts are spontaneously created whenever we confront a novel multiple input–output environment or whether these concepts are discovered only by exceptional individuals and then later are incorporated into our general world knowledge through exposition. This general issue was also raised by Qin and Simon (1990) with regard to the discovery of scientific laws.

There are several reasons it might be likely that intervening concepts are spontaneously created. One reason is that it simplifies the learning problem. As we pointed out earlier, the use of intervening concepts reduces the number of input–output relations that need to be learned (compare the top and bottom panels of Fig. 1). Perhaps a more important reason is that new relationships can be directly inferred from existing relationships. For example, suppose

you have learned that an intervening concept (e.g., worker ability) mediates between two inputs (e.g., education and experience) and two outputs (e.g., worker productivity and worker efficiency). Further, you have learned that the relations between the inputs and the intervening concept are positive, as are the intervening concept-to-output relations. Suppose a new output is considered (e.g., worker absenteeism). If worker absenteeism is observed to be negatively related to worker productivity, then one could use the intervening concept to infer that it will also be negatively related to the other output variable and negatively related to all of the input variables. Of course, as with any theoretical inference, there is a danger that these inferences will be incorrect. Nevertheless, the fact that intervening concepts support inferences reflects the function of causal concepts in general.

There are also several reasons people may not invariably form intervening concepts when they encounter a novel input–output environment. One reason is that abstracting far beyond the immediate concrete situation presumably requires considerable thoughtful effort. Forming an intervening concept is not necessary to achieve an optimal level of performance in predicting the outputs from the inputs to a causal system. The brute force use of direct input–output relations can in principle achieve the same level of performance (this is illustrated in detail in Experiment 1), and in currently popular learning models (adaptive network), learning might well be as easy with a larger number of links (input–output relations) as with a small number of links. Another reason is that in some situations the use of an intervening concept may provide an oversimplified representation of the actual situation. For example, if each output were independently related to the inputs (as suggested by the top panel in Fig. 1), then the use of an intervening concept would produce suboptimal performance and incorrect inferences.

The above considerations lead to two fairly extreme views about the construction of intervening concepts. One view is that subjects will always spontaneously create these concepts in order to simplify a relatively complex and novel input–output environment. The second view is that subjects will rarely or even never form such concepts because of the thoughtful effort required. However, both of these extremes leave out the potential influence that the actual causal environment has on the formation of intervening concepts. Causal environments vary in the extent to which they afford the formation of intervening concepts. Subjects may be able to discriminate between (a) causal environments in which all the outputs are actually determined by a single common factor, and (b) causal environments in which each output is an independent function of the inputs. This possibility leads to the intermediate viewpoint that subjects are sensitive to the regularities generated by an environment in which an intervening variable is operative, and they form such concepts only when these concepts are afforded by the environment. These hypotheses are tested in the first two experiments.

Given that there are subjects and situations for which intervening concepts

are formed, another issue concerns the degree to which the learner continues to rely on the intervening concept to draw inferences when new variables are added to the environment. If subjects have acquired a coherent understanding of the intervening concept, then we would expect that they will apply the concept for deriving inferences concerning newly introduced relationships. Alternatively, if subjects do not demonstrate inferences drawn from the intervening concept, then it would raise questions concerning the depth of the subject's understanding of the concept. This issue is addressed in the third and fourth experiments.

In sum, the following experiments provide three different converging lines of evidence (cf. Garner, Hake, & Erickson, 1956) for the learning of intervening concepts. The first is based on a principal component analysis of subjects' output responses, the second is based on the accuracy of subjects' inferences to a new fourth output, and the third is based on the accuracy of subjects' inferences from a new fourth input.

Finally, how are intervening concepts learned? There are at least two existing theoretical approaches to concept learning that are relevant. One approach rests on the obvious analogy between intervening concepts and hidden-unit adaptive network learning models. Both are used to recode inputs into a hidden variable (unit), and this hidden variable then mediates the relations between inputs and outputs. Exemplar-based learning models provide a completely different view of concept learning. Given their success in the category learning domain (Nosofsky & Kruschke, 1992), these models merit consideration in more complex domains. To foreshadow, neither of these two approaches as instantiated currently can account for the learning evidenced in the following experiments. After reporting the results, we develop a network model in which learning is guided by parsimony constraints, and we demonstrate that in acquiring intervening concepts the model is sensitive to the environment, as are our human learners.

EXPERIMENT 1

In our basic paradigm, subjects were exposed to fictional manufacturing systems that had five inputs (raw materials) and three outputs (manufactured products). The subjects' task was to predict the amount of each output product that would be produced by a system for a given input pattern. Subjects were given 100 trials of training with each of 10 factory systems (1 system per daily session), where each system differed according to the weights of the links connecting the inputs to the outputs. Although the weights differed from system to system, all 10 systems shared the same causal structure for any particular subject (the type shown in either the top or the bottom panels of Fig. 1).

Consider now what subjects might learn if they were exposed only to the environmental input–output relations (i.e., exposed to many pairings of input and output values). If these input–output relations were actually mediated by an unknown intervening factor as schematized in the bottom panel of Fig. 1,

then subjects might eventually discover this and formulate an intervening concept. That is, computationally speaking, subjects would be learning that the five input values are combined additively in a weighted fashion to produce an intermediate value, and this intermediate value in turn is moderated by each of three weights to yield the level (value) of each of the three outputs (for the given set of inputs).

On the other hand, subjects could simply learn individual input–output relations. That is, the intervening-concept structure just mentioned can be represented in a mathematically equivalent form as a set of 15 input–output relations (as shown in the top panel of Fig. 1). Accordingly, subjects could be learning that each input has a unique weighted relation to each output, and that deriving the level (value) for a particular output involves combining additively the five unique weighted inputs associated with a particular output. If this were so, then subjects would not have discovered the intervening variable. It is critical to note that learning either structure (the intervening-concept structure or the input–output structure) would achieve equally accurate performance as measured by the match between the predicted and correct outputs. To see this, assume that the to-be-learned intervening-concept system has the particular set of weights displayed in Table 1 (associated with the diagram in the bottom panel of Fig. 1). Translating this system into an input–output system (i.e., multiplying each input-to-intervening variable weight with each intervening-variable-to-output weight) yields another weight set, also displayed in Table 1. Now, assume that the subjects receive the five input values, 1.0, 2.0, 3.0, 4.0, and 5.0. Applying the weights from the intervening-concept system yields 2.25 for output 1, −1.45 for output 2, and 4.10 for output 3. Applying the weights from the input–output weight set produces exactly the same output values just listed. Therefore, performance accuracy per se is not discriminating in terms of whether an intervening concept or a set of individual input–output relations has been learned. Accordingly, the main objective of Experiment 1 was to establish analytic techniques for discriminating between these two possibilities.

In this training situation, we suggest the following major indicator for revealing acquisition of an intervening concept. The proposed indicator involves a principal component analysis (see Tatsuoka, 1988) of the three output predictions produced by each subject. This analysis computes three factors, and each output-response measure can be expressed as a linear combination of the three factors. The first factor (first principal component) is constructed to account for the largest amount of variance in the three output-response measures. For a system in which the environmental inputs and outputs are mediated by an intervening concept, as subjects' predictions become more accurate (e.g., near the end of training) the first factor will necessarily reproduce more and more of a subject's output predictions (i.e., the percentage of variance in the three output responses reproduced by the first principal component should be approaching 100%). This will occur regardless of whether the

TABLE 1
Weight Structures Supporting Equivalent Performance in an Environment with an Intervening Variable

Intervening-variable weight structure

Inputs to intervening variable		Intervening-variable to outputs		
Input No.		Output No.		
(1)	.68			
(2)	.30	(1) .45	(2) −.29	(3) .82
(3)	.16			
(4)	.76			
(5)	.04			

Input–output translation of above structure

	Inputs to outputs		
	(1)	(2)	(3)
(1)	.31	−.20	.56
(2)	.14	−.09	.25
(3)	.07	−.05	.13
(4)	.34	−.22	.62
(5)	.02	−.01	.03

Note. Each weight in a cell of the bottom table is obtained by multiplying the corresponding row and column weight in the top tables.

subject has learned an intervening concept or a set of input–output weights (because in such a system the outputs will be correlated, so that learning the outputs will result in a correlated set of predictions). Therefore, after some training on a system, the principal component analysis will not be telling in terms of the underlying representation acquired. The critical indicator for acquisition of an intervening concept will be if the percentage of variance reproduced by the first principal component (for the three output responses) approaches 100% *at the beginning of training on each new system.*[1] After forming an intervening concept, all subsequent predictions of outputs can be based on a single intervening variable, even though the specific weights for each new causal system are initially unknown. In light of these considerations, in the first three experiments, the principal component analysis was always conducted at the first block of trials for each training system.

[1] If subjects are using an intervening concept to form their predictions at the beginning of training on each new causal system, then the principal component is expected to account for nearly, but not exactly, 100% of the output variance for two reasons. First, there will be some response error. Second, the weights connecting the inputs to the intervening factor are still changing across trials at the beginning of training.

In order to test the validity of the first principal component measure of intervening-concept learning, we attempted to strongly encourage acquisition of an intervening concept for some subjects, and to strongly encourage acquisition of individual input–output relations for other subjects. This was accomplished by training one group of subjects on a causal system with an *actual intervening variable* (like the bottom panel of Fig. 1) and another group on an *input–output* system that had no single intervening factor (like the top panel of Fig. 1). Thus, the latter group could not accurately perform the task by constructing a single intervening concept. This group was included to gauge the pattern of the principal component measure for subjects for whom it could be reasonably assumed that no intervening concept was acquired. Significant deviations from this baseline on the principal component measure could then be taken as evidence that the intervening-concept group did form an intervening concept (rather than input–output associations). To *initially bias* subjects toward the appropriate causal structure, at the outset of training subjects were shown a diagram of the type of structure underlying the input–output relations (similar to Fig. 1, top or bottom, depending on the group).

The critical prediction is that the percentage of variance reproduced by the first principal component should be larger for the intervening-variable group than for the input–output group, and this difference should increase across training sessions with the 10 causal systems as intervening-variable subjects become more disposed/experienced with using an intervening concept. It is worth mentioning again that the accuracy of subjects' predictions may or may not differ across groups; either pattern could obtain even in the presence of differences in what is learned (intervening concept vs individual input–output relations).

To corroborate the veracity of the principal component analysis, we added special features to the present experiment to allow converging tests of whether or not the intervening concept was learned (in the intervening-variable group). For the intervening-variable group, the weights linking the intervening variable to the outputs were held constant across all 10 causal systems (although the weights connecting the inputs to the intervening variable changed across systems). If subjects in this group formed an intervening concept, then they should be sensitive to the constant relations between the outputs, even though the input–output relations were changing across systems (due to changes in the weights connecting the inputs to the intervening factor). This would be reflected in significant correlations between each of the three possible pairings of the subjects' three output predictions. Specifically, if subjects formed the intervening concept in the intervening-variable group, then the correlation between the first and third output should be positive, and the correlations between the second output and the other two outputs should be negative. This is because the weights linking the intervening factor to the first and third outputs were positive, and the weight linking the intervening factor to the

second output was negative. Importantly, the pattern of correlations just described should be evident even at the beginning of training on each causal system, once the concept has been acquired.

In contrast, for the input–output group, the correlations should be near zero. Although exactly the *same* set of weights used in the intervening-concept systems were used to construct the input–output weights (details given under Methods), the assignment of these weights to input–output relations was randomly rearranged for each of the 10 systems. Thus, in the input–output condition there was no uniform relation between the outputs across the systems for subjects to exploit. If, however, for whatever reason correlations between predictions of the three outputs are simply due to some response bias, then high correlations would appear in the input–output group as well.

Method

Subjects and design. The subjects were 12 volunteer undergraduate or graduate students who were paid between $3.50 and $5.50 per hour, according to their performance. One subject dropped out during the experiment, leaving 11 subjects. These subjects were assigned to two experimental groups. One group of 6 subjects (the intervening-variable group) received multiple input–output systems constructed so that the inputs were associated with the outputs via a ''hidden'' intervening factor (similar to that shown at the bottom of Fig. 1). The 5 remaining subjects (the input–output group) received multiple input–output systems constructed without an intervening factor (similar to that shown at the top of Fig. 1).

Procedure. The instructions were printed in a three-page booklet (one booklet for each experimental group), which was given to subjects at the beginning of the experiment. Subjects were asked to imagine that they were learning how to operate the controls of an industrial factory system in which there were five raw materials (five inputs) that yielded three different products (three outputs). The subjects were further provided a description and diagram of the kind of system they would encounter. Accordingly, the intervening-variable group saw a diagram of a causal system with an intervening variable (two subjects—Subjects 3 and 4—were inadvertently not shown a causal diagram), and the input–output group received a diagram of an appropriate input–output structure (with no intervening variable). (The diagrams were essentially those presented in Fig. 1, but with no labels.)

Subjects were told that at the beginning of each trial the amounts of the five different materials would be displayed as vertical bars on the upper half of a video screen. They were informed that each vertical bar would represent one material (the materials were left unnamed), and that the height of each bar would represent the amount of the material. Subjects were instructed that they would have to try to predict the amount of the

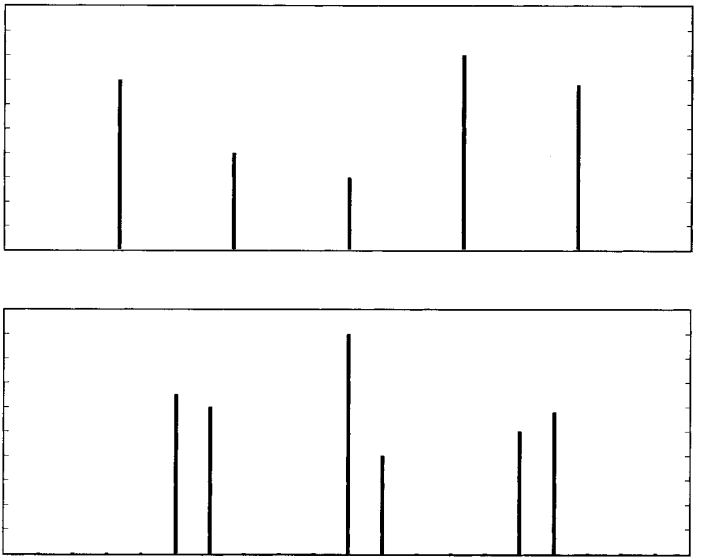


FIG. 2. A typical computer display of the input, a subject's prediction, and the correct amount of the output. The top panel represents inputs. The left bar within each pair of adjacent bars represents a subject's prediction and the right bar represents the correct amount of the output.

three output products manufactured by the system on the basis of the displayed input material. To prevent discouragement, subjects were forewarned that the task was very difficult. They were cautioned to try to be as careful as possible in entering their predictions, and that their pay would be based on the accuracy of their predictions. Prior to training, subjects were given one example trial to familiarize them with the procedure. For each trial, subjects made their prediction by adjusting the height of three vertical bars generated on the lower half of the video screen. None of the bars was initially displayed in this half of the display; subjects generated the bars by manipulating four arrow keys. After subjects drew their predictions, the correct amount of each output product produced in that factory was drawn beside the subject's corresponding drawing of each output product.

Figure 2 shows a typical computer display of input, the subject's prediction, and the correct amount produced. The first bar of each pair of bars in the lower box of the screen represents the subject's prediction, and the second bar represents the correct amount produced. Finally, on each trial subjects were given a performance evaluation based on the sum of absolute deviations between their predictions and the correct amount produced. The performance evaluation ranged from 0/100 to 100/100, where 0/100 represented the worst performance and 100/100 represented best performance.

Each subject received 100 trials of training with each of 10 different factory systems. Each system took about one hour, and one system was presented during each of 10 consecutive weekdays.

Stimuli. The 100 trials for each intervening-variable factory system were constructed as follows. Each of the five input weights linking an input to the intervening factor was randomly sampled from a uniform [0, 1] distribution. The same set of five input weights was used for all 100 trials of a given factory system (within the same session), but a different set of five input weights was sampled for each new factory system (a new session). The three output weights connecting the intervening variable to the three outputs were held constant across all trials and systems (0.8, −1.0, and 0.6, for outputs labeled 1, 2, and 3, respectively). On each trial, five input values were sampled from a uniform [0, 50] distribution. The value of the intervening variable was computed by summing the five products formed by multiplying each input value by the corresponding input weight. Each of the three output values was generated by multiplying the value of the intervening variable by one of the three output weights.

The 100 trials for each input–output factory system were constructed from the weights and input values of a corresponding intervening-variable factory system as follows. Using the weights from one of the intervening-variable factory systems, each of the 5 input weights was multiplied by each of the 3 output weights to form $5 \times 3 = 15$ input–output weights. Then each of these 15 input–output weights was randomly assigned to one of the 15 input–output relations to form an input–output factory system. For each trial, the 5 input values sampled for the intervening-variable factory system were used again as the input values for the corresponding input–output factory system. Finally, each of the 3 outputs for the input–output factory system was calculated by summing the 5 products formed by multiplying each input value by the corresponding input–output weight for a particular output.

Results

The results are presented in three parts. First, we examined whether subjects improved the accuracy of their predictions as training proceeded; that is, did learning occur? The second and most important part presents the results of the principal component analysis, which is our main measure of intervening-concept learning. The last part presents the correlations among the three output responses, which provides another assessment of the sensitivity of subjects to the hidden structure for the intervening-variable training group.

Prediction accuracy. Prediction accuracy was assessed by two related measures. One was the absolute prediction error between each of the subject's predictions and the corresponding correct output. These absolute prediction errors were first averaged across the three responses and then averaged across blocks of 25 trials, producing a single mean absolute error index for each subject and training block. The second measure of accuracy was the correla-

TABLE 2
Mean Prediction Accuracy (mean Absolute Prediction Error and Correlation between Predicted and Actual Outputs) for Experiment 1

		Learning group	
Accuracy measure	Trial block	Intervening-variable	Input–output
MAE	1	6.21	4.72
	2	5.23	4.06
	3	4.97	3.87
	4	4.87	3.96
Correlation	1	0.44	0.10
	2	0.53	0.20
	3	0.55	0.21
	4	0.58	0.20

Note. Each trial block represents 25 trials. MAE, mean absolute error.

tion between each of the subject's predictions and the corresponding correct output for a block of 25 trials. These three correlations were averaged across the three outputs to produce a single correlation index for each subject and training block. The main difference between these two accuracy indices is that the correlation index is only sensitive to the match between the predicted and correct response patterns, whereas the mean absolute error is also affected by the mean and standard deviation of the subject's predictions.

Table 2 shows the basic results for each group and accuracy index, separately for each quarter (25 trials) of training. As can be seen from this table, both groups produced clear improvement in predictive accuracy on both measures. Separate repeated-measures analyses of variance (ANOVA) for each group on each measure confirmed that the observed improvements were significant (for absolute prediction error: $F(3,15) = 18.76$, $MSe = 0.12$, $p < .0001$, and $F(3,12) = 11.83$, $MSe = 0.06$, $p = .0007$ for the intervening-variable and input–output groups, respectively; for the correlation indices: $F(3,15) = 9.71$, $MSe = 0.002$, $p = .0008$, and $F(3,12) = 6.09$, $MSe = 0.003$, $p = .0093$ for the intervening-variable and input–output groups).

It is tempting to compare the predictive accuracies across groups, but this comparison is complicated for the following reason. Both accuracy measures are defined in terms of the correct output responses, which were not equated across groups. Thus, the differences between groups may be due to the differences in the input–output pairings programmed for each group and may not necessarily be due to differences resulting from differential use of an intervening concept.[2]

[2] While it is difficult to compare across groups, one feature is worth a brief mention. The input–output group produced a lower mean absolute error than the intervening variable group, but

Principal component analysis. The percentage of variance in the subjects' output responses reproduced by the first principal component was computed separately for each subject and causal system as follows. First, a 25×3 data matrix was formed from the first 25 trials of a causal system, and for the 3 output predictions produced by a subject. Then a 3×3 correlation matrix was computed by correlating each pair of columns of the 25×3 matrix. The first eigenvalue was abstracted from this correlation matrix, and, finally, the proportion of variance in the three outputs reproduced by the first principal component was computed by dividing the first eigenvalue by the sum of all three eigenvalues (see Tatsuoka, 1988).

Table 3 shows the percentage of variance reproduced by the first principal component for each subject and system. Inspection of this table shows that for half of the subjects in the intervening variable group, the first principal component accounted for over 90% of the variance for at least six of the systems. This pattern was not evident for any of the subjects in the input–output condition. Further, for all but one of the subjects in the intervening-concept condition, on average a substantial amount of variance (over 70%) was reproduced by the first principal component, whereas this was the case for only one of the subjects in the input–output condition.

A two-factor mixed ANOVA on these data (with group and system as factors) confirmed that the percentage of variance reproduced by the first principal component was significantly larger in the intervening-concept condition than in the input–output condition, $F(1,9) = 5.49$, $MSe = 0.17$, $p = .0438$. Importantly, this effect significantly interacted with system, $F(9,81) = 2.66$, $MSe = 0.006$, $p = .0094$. Planned comparisons on the first and last system showed there was no difference between the groups with subjects' first experience with the systems ($F < 1$). As subjects gained more experience with the systems, robust differences emerged between the two conditions, and the comparison on the last system was significant, $F(1,81) = 34.08$, $MSe = 0.006$. There was also a significant main effect of system, $F(9,81) = 6.38$, $MSe = 0.006$.

Correlations between each pair of output responses. For reasons outlined in the introduction, we calculated the correlations between each pair of subjects' predictions for the first 50 trials of each system. Table 4 displays these correlations. As expected for the intervening-variable group, after some experience with the systems, the correlations were consistently large and positive

---

the opposite occurred with the correlation measure. The explanation is simple but uninteresting. Essentially, the input–output group concentrated on learning the appropriate mean output, independent of the input pattern. The mean output has a strong influence on the mean absolute error measure, but no effect on the correlation measure. The latter measure is only sensitive to the linear relationship between subjects' outputs and the correct outputs, it is insensitive to the subjects' mean output and the correct mean output. Furthermore, the order of MAE was reversed in the next three experiments (i.e., the intervening group produced a smaller MAE).

TABLE 3
Proportion of Variance Reproduced by the First Principal Component for Each Subject on the First 25 Trials of Each System (Experiment 1)

	Intervening-variable						
System	S3	S4	S42	S43	S44	S47	Mean
1	.62	.53	.64	.64	.49	.55	.58
2	.87	.63	.59	.77	.46	.81	.69
3	.87	.66	.86	.86	.51	.95	.79
4	.92	.91	.85	.95	.43	.94	.83
5	.78	.68	.86	.83	.39	.86	.73
6	.95	.68	.87	.94	.41	.92	.80
7	.95	.67	.71	.94	.43	.96	.78
8	.93	.71	.88	.95	.60	.97	.84
9	.97	.77	.94	.98	.55	.97	.86
10	.95	.82	.67	.93	.58	.91	.81
Mean	.88	.71	.79	.88	.49	.88	

	Input–output					
System	S31	S32	S34	S35	S36	Mean
1	.44	.72	.49	.44	.56	.53
2	.52	.61	.56	.42	.80	.58
3	.74	.68	.62	.49	.80	.67
4	.66	.63	.69	.51	.86	.67
5	.70	.59	.53	.45	.48	.55
6	.66	.63	.59	.39	.91	.64
7	.62	.50	.43	.43	.59	.51
8	.48	.55	.53	.52	.75	.57
9	.66	.68	.55	.59	.67	.63
10	.49	.45	.42	.59	.67	.52
Mean	.60	.60	.54	.48	.71	

between the first and third outputs, and consistently large and negative between the other two pairings. In contrast, for the input–output group there was no systematicity in the magnitude and direction of the correlations for any output pairs, and generally the correlations were low.

Discussion

Considering the complexity of the multiple input–output learning task used in this experiment, it was initially unclear whether subjects would be able to learn to improve their prediction accuracy during the relatively limited amount of training that they received with each causal system. Nevertheless, subjects did exhibit a substantial amount of learning. Significant improvements in predictive accuracy occurred for both the intervening-concept group and the input–output group. Further, informal comparisons of the predictive accuracy across the two groups (see Table 2 for means) indicates that the input–output

TABLE 4
Correlations between Each Possible Pair of the Subjects' Predictions Based on the First 50 Trials for Each System (Experiment 1)

	Intervening-variable			Input−output		
System	Outputs 1,3	Outputs 1,2	Outputs 2,3	Outputs 1,3	Outputs 1,2	Outputs 2,3
1	0.30	0.03	0.03	0.38	−0.16	−0.02
2	0.72	−0.30	−0.17	−0.10	0.24	−0.02
3	0.75	−0.65	−0.49	−0.34	−0.10	0.60
4	0.86	−0.74	−0.71	−0.42	0.59	0.16
5	0.79	−0.61	−0.46	−0.02	0.12	0.25
6	0.76	−0.67	−0.61	0.36	0.30	0.35
7	0.80	−0.65	−0.54	0.38	0.15	0.18
8	0.82	−0.77	−0.68	0.41	−0.24	0.35
9	0.79	−0.79	−0.76	0.10	0.16	0.12
10	0.75	−0.63	−0.54	0.25	0.10	0.30

group, if anything, tended to generate more accurate predictions than the intervening-variable group. Thus, as indicated earlier, the accuracy data leave unclear whether the intervening-variable group extracted an intervening factor or instead formed direct input−output associations.

To answer this main question one must examine the principal component results. The first principal component reproduced most of the variance in the output responses for the subjects in the intervening-variable group, whereas this was not the case for the subjects in the input−output group (with one exception in each group). Importantly, this was true for the first trial block of a causal system, before subjects could even learn the appropriate weights of the new system. Furthermore, this difference between the two groups increased across training sessions with the 10 causal systems. No differences between groups appeared on the first few causal systems. As exposure to the causal structure increased, the first principal component increased for the intervening-variable group but not for the input−output group. This pattern is telling because it demonstrates that when input−output associations are mediating responses (the input−output group), the first principal component does *not* account for a substantial amount of variance in the set of predictions generated by subjects. Accordingly, the implication is that subjects in the intervening-variable group, for whom the first principal component accounted for most of the variance in their predictions, were using an intervening factor to formulate their predictions.

The conclusions from the principal component analysis were corroborated by the correlations between each of the three pairs of the subjects' output responses from the first two trial blocks. Recall that output weights were

fixed across systems to .8, −1.0, and .6 for outputs 1, 2, and 3, respectively. Nevertheless, if one only learned input–output relations, then this invariance would not be noticed at the beginning of training for each system because the 15 observed input–output relations changed from system to system for the intervening variable group. The key result from these correlations concerns the signs of the correlations. Note that the correlations for the intervening-concept group are high and match the direction of the relationship between each pairing of correct outputs (produced by the fixed output weights), but this pattern occurs only after the first system. In addition, this pattern of results did not occur for the input–output group, for whom there was no underlying invariance across systems.

These results lay an important foundation for further experimentation into intervening-concept learning by establishing analytic techniques for revealing concept formation. Specifically, the usefulness of principal component analysis was verified. When the percentage of variance reproduced by the first principal component is high (i. e., near 100), then the implication is that performance is primarily mediated by one underlying factor. This assumption was supported in the present domain. In contrast, enhancement of prediction accuracy did not signal performance mediated by an underlying factor, at least in the present context.

To establish the methodology for this kind of learning task, we purposely designed the environment to make the intervening concept salient by giving subjects explicit instructions concerning the presence or absence of an intervening variable. Interestingly, despite these instructions, no differences between the groups in intervening-concept learning appeared in the first initial causal systems. On the one hand, this finding may reflect the inability of subjects to make effective use of the prior instructions concerning the causal structure for each group. On the other hand, the effect of instructions may not appear at an early stage, but may be crucial for the learning process that eventuates in the acquisition of the intervening concept. In the next experiment we examined this issue by eliminating the prior instructions about the causal structure. In addition, we continued our investigation of intervening-concept learning by implementing a new cover story for the causal system to extend the generality of this research.

EXPERIMENT 2

One objective of this experiment was to demonstrate that the concept learning obtained in Experiment 1 was not limited to one idiosyncratic context circumscribed by the specific features chosen for the first experiment. Accordingly, the input–output environments were presented under a different cover story in which subjects were told that they would be learning about a biological cell system rather than a factory system. The cell system had three critical features (cell-wall thickness, cell height, and density of interior substance)

that could take on different values, and subjects were asked to predict the amounts of three different products produced by this cell system.

Another objective was to investigate the extent to which subjects form intervening concepts in the absence of explicit instructions regarding the structure of the input–output environment. One group of subjects received training with a causal system that contained an intervening variable, and another group received training with an input–output structure with each output an independent function of the inputs. Logically, four hypotheses can be constructed by considering a combination of two factors—subjects' initial bias and their sensitivity to the training environment. The first hypothesis is that very few subjects ever discover the intervening factor, and the majority of subjects simply learn all of the input–output relations separately in a brute-force manner. A second hypothesis is that all subjects start out from the very beginning using an intervening concept to simplify the learning problem, and they persist in using this concept throughout training, independent of the actual causal structure. The first hypothesis implies that the first principal component will not account for most of the variance in the subjects' output predictions, whereas the second hypothesis implies that most of the variance will be reproduced by the first principal component throughout the training period.[3] Both of these hypotheses imply that there will be no effect of group (on the principal component measure).

In contrast, the last two hypotheses assert that subjects are sensitive to the causal structure of the environment, so that whether they eventually form an intervening concept depends on the nature of the causal environment. This implies that there will be group differences, and furthermore these differences will increase as the subjects receive more experience with the causal structure. The third hypothesis asserts that subjects do not start out with an intervening concept, but they will form a concept after exposure to an environment that has an intervening factor. The last hypothesis asserts that subjects start out imposing an intervening concept on all environments (see Footnote 3), but then they discard this concept when confronted with an environment without an intervening factor.

Related to these last two hypotheses is the possibility that subjects are influenced by the format in which the inputs are presented. Some input formats may influence the salience of the presence or absence of an intervening factor in the system. Specifically, inputs presented in an integral fashion might suggest the presence of an underlying single factor more so than inputs presented in a sequential, separate manner. Accordingly, the particular inputs were presented integrally for some subjects, and separably for other subjects.

[3] Hypotheses 2 and 4 were not supported by the first experiment, but they are worth reconsidering in the second experiment because of the change in cover story and the lack of any prior instructions about the type of causal system.

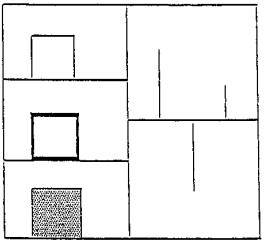
If this kind of influence holds, then we would expect the integral format to enhance the learning of the intervening concept. Before we present the experiment, it is worth noting that the computation and memory demands of the task were simplified so that less training would be needed. The changes in this regard were (a) reducing the number of input variables from five to three; (b) restricting the number of input values to four equally spaced values; (c) increasing the range of output values to increase the impact of the response pattern on the mean absolute error performance measure; and (d) during feedback, reshowing subjects the inputs and correct outputs from the previous trial. Unlike in the previous experiment in which the output weights were held fixed for the intervening-variable group, none of the weights were held constant across systems in this experiment.[4]

Method

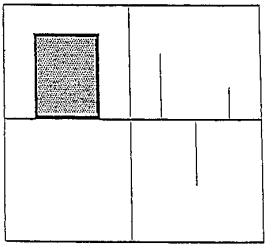
Subjects and design. Twenty-seven volunteers, paid as in Experiment 1, served as subjects. The subjects, none of whom participated in Experiment 1, were undergraduate students enrolled at Purdue University. The subjects were randomly assigned to one of four conditions defined by the factorial manipulation of two variables, each with two levels. One variable represented the structure of the input–output systems presented to the subjects, with some subjects receiving systems with an *intervening factor* and other subjects receiving systems in which there was no intervening factor (labeled the *input–output* condition). The other variable represented the way in which the inputs were presented. In the *integral* presentation format the three input dimensions were presented in one integrated pattern (a schematic of a cell), and in the *separable* presentation format the input dimensions were pictured separately (Fig. 3 provides example displays). There were six subjects in the input–output separable group and seven in each of the other groups. There were four other subjects tested who elected not to complete the experiment (three in the input–output separable condition and one in the input–output integral condition).

Procedure. All subjects were given the same cover story: ''The present experiment simulates a hypothetical cellular system in which the levels of the cell's chemical outputs are determined by its physical characteristics. Each cell differs along three important dimensions: the height of the cell (the cells are rectangular in shape); the thickness of the cell wall; and the density of the cytoplasm inside the cell.'' Subjects were then referred to a figure illustrating how the ''cell'' would appear on the computer monitor, with the particular figure appropriately matched to the integral and separable conditions

[4] The fact that the output weights changed across systems in the second experiment implies that we do not expect the sign of the correlations between each pair of output predictions to remain constant across systems in the second experiment.



Separable input condition



Integral input condition

FIG. 3. An example display of separable and integral input conditions. The top panel represents a separable input condition and the bottom panel represents an integral input condition. Bars in the right-hand side of each panel represent a subject's prediction.

(the variation in the figure was the only difference in the instructions across subjects). The subjects were informed that there would be four different possible values for each dimension and that their task was to predict the amounts of the outputs based on the input levels (the cell characteristics). Subjects were not given any instruction or diagrams alerting them to the structure of the system (e.g., the presence of an intervening factor).

The presentation of the input values on each trial was accomplished by a graphic representation of the input dimensions. Thus, cell-wall width was presented with lines of differing thickness, cell height was represented by differing heights, and cytoplasm density was represented by relative density of dark specks on a light background (see Fig. 3). (The zero-input value was represented by some minimal thickness, height, or density.) As in Experiment 1, subjects were not given a numerical presentation of the input values—only the graphical presentation was used. The procedure by which subjects produced their predictions and the presentation of the correct outputs were identical to that in Experiment 1. The one additional feature was that after displaying the correct output on the current trial, the display (without blank-

ing) presented the inputs and outputs on the preceding trial as well. To conclude the trial, the evaluation score was presented. The prediction-accuracy evaluations in this experiment (based on the mean absolute error between the predictions and the correct outputs) were submitted to a linear transformation designed to lower the scores. The intent here was to make it somewhat more difficult for subjects to achieve accuracy scores near 100 (our feeling was that high scores too early in training might produce some complacency).

After the instructions, subjects were given five practice trials on an example system to familiarize them with the display and the response procedure. They then were given the 50 trials for the first system, and dismissed for the day. On the next two successive days subjects returned to complete the next four systems (two systems per day, with 50 trials per system). One subject in the intervening-factor condition with separable stimuli did not return for the third day, and upon being contacted, this subject returned to complete the third session two days later.

Stimuli. Subjects received five systems, and all systems contained three inputs and three outputs. In this experiment, the input values sampled were always restricted (for all three inputs and all five systems) to four equally spaced values (0, 33, 66, 99). The weights relating inputs to outputs, however, changed across the systems for both the intervening factor and the input–output groups. Thus, unlike in Experiment 1, both the input weights linking each input to the intervening factor and the output weights linking the intervening factor to the outputs changed from system to system. Second, the weights were sampled from the interval $[-1, +1]$, so that both positive and negative weights were allowed for both the input and output weights of an intervening factor system. The remaining details for generating stimuli are the same as in Experiment 1.

Results

The results are organized into two parts. The first part presents an analysis of predictive accuracy as a function of training block. The second presents the more critical principal-component analyses.

Predictive accuracy. The same two indices of predictive accuracy described in the first experiment were used again for this experiment—the mean absolute prediction error, and the correlation between the subject's prediction and the correct output, computed for each block of 25 trials.

Table 5 shows the basic results for each group and accuracy index, separately for each trial block of training, averaged across systems and subjects. Similar to the first experiment, both the intervening-variable groups and the input–output groups showed improvements across training trials for both accuracy measures. Two-factor mixed ANOVAs (with trial block as a within-subjects variable and input format as a between-subjects variable) indicated that the trial-block improvements were significant for mean absolute prediction error ($F(1,12) = 37.48$, $MSe = 0.62$, $p < .0001$, and $F(1,11) = 243.69$,

TABLE 5
Mean Prediction Accuracy (Mean Absolute Prediction Error and Correlation between Predicted and Actual Outputs) for Experiment 2

		Intervening-variable		Input–output	
Accuracy measure	Trial block	Integral	Separable	Integral	Separable
MAE	1	14.51	15.80	20.94	21.93
	2	12.74	13.92	16.38	17.20
Correlation	1	0.30	0.23	0.17	0.10
	2	0.48	0.42	0.30	0.22

Note. Each trial block represents 25 trials. Integral and separable refer to stimulus format (see text for details).

$MSe = 0.57$, $p < .0001$ for the intervening-variable and input–output groups) and for the correlational measure ($F(1,12) = 40.83$, $MSe = 0.006$, $p < .0001$, and $F(1,11) = 75.84$, $MSe = 0.001$, $p < .0001$, for the intervening-variable and input–output groups). Input format failed to show any systematic influences on prediction accuracy, either in terms of main effects (largest $F = 1.73$) or in terms of interactions with trial block (F's < 1).

Another correlational analysis was performed to check the following simple hypothesis: Perhaps subjects carry over input–output associations learned at the end of training from the old system (n) to the initial trials of the new system ($n + 1$). To test this notion, first we computed a correlation between the subjects' average outputs to the last 10 stimuli on system (n) and the correct outputs to these same stimuli from the same system (n). The average correlations were $r = .88$ and $r = .83$ for the intervening and input–output groups, respectively, averaged across systems and output responses. Next we computed the correlations between the subjects' average outputs produced by the first 10 stimuli at the beginning of a new system ($n + 1$) and the ''correct'' outputs produced by the old system (n) to these same 10 stimuli. The average correlations in this case were only $r = .09$ and $r = .03$ for the intervening and input–output groups, respectively. This provides evidence against the idea that subjects carry over specific input–output associations from one system to the next.

A final correlational analysis was performed to check another hypothesis: Perhaps subjects learn to focus on only one input variable and ignore the other two inputs. If this hypothesis is correct, then only one of the three inputs should be significantly correlated with all three of a subject's outputs after extensive training. This hypothesis was tested separately for each subject by regressing the three inputs on the outputs from the last 10 trials of training on the last system. For the intervening variable group, of the 11 subjects that yielded at least one significant input effect, 8 produced more than one signifi-

TABLE 6
Proportion Variance Reproduced by the First Principal Component for the First 25 Trials of Each System (Experiment 2)

	Intervening		Intput–Output	
System	Integral	Separable	Integral	Separable
1	0.78	0.65	0.48	0.51
2	0.84	0.72	0.50	0.57
3	0.93	0.79	0.54	0.61
4	0.93	0.87	0.48	0.51
5	0.88	0.76	0.52	0.52

cant input. For the input–output group, of the 11 subjects that yielded at least one significant input effect, 10 produced more than one significant input. In conclusion, a great majority of subjects were influenced by more than one input variable.

Principal component analysis. The percentage of variance in each subject's output predictions reproduced by the first principal component was computed from the first 25 trials of each system. In general, the first principal component accounted for most of the variance in the intervening-factor groups' predictions (82%) but not in the input–output groups' predictions (52%). A three-way mixed ANOVA revealed that this difference was significant, $F(1,23) = 160.65$, $MSe = 0.018$, $p < .0001$. This effect significantly interacted with system, $F(4,92) = 5.00$, $MSe = 0.007$, $p = .0011$, further revealing that predictions for the intervening variable groups were increasingly reproduced by one principal component, whereas this trend did not emerge for the input–output groups (see Table 6 for the means). Planned comparisons (contrasting intervening-variable vs. input–output subjects) for the first and last systems within each input–format condition (integrated or separable) showed that one principal component reproduced more of the variance in the intervening-variable group's predictions (relative to input–output subjects) even by the first system for integral stimuli, $F(1,92) = 47.37$, and for separable stimuli, $F(1,92) = 9.33$, $MSe = 0.007$. On the last system, the differences remained significant, $F(1,92) = 68.21$ and $F(1,92) = 27.43$, $MSe = 0.007$ for integral and separable stimuli, respectively.

Finally, input format also significantly interacted with experimental group (intervening variable, input–output), $F(1,23) = 11.23$, $MSe = 0.018$, $p = .0028$. Inspection of Table 6 shows that integral inputs increased the percentage of variance in the predictions reproduced by the first principal component for the intervening variable subjects (relative to the separable input form), but not for the input–output subjects.

In addition, the percentage of variance in each subject's output predictions

reproduced by the first principal component was computed from the very first trial of systems 3, 4, and 5 (before any feedback). The amount of variance reproduced by the first principal component increased across systems for the intervening variable condition (74, 64, and 90%, respectively for the third, fourth, and fifth systems). On the other hand, the amount explained by the first principal component remained constant across systems for the input–output condition (45, 50, and 44%, respectively for the third, fourth, and fifth system). Thus, the basic results were replicated using only the very first trial, before any learning about the specific input–output relations for a system could occur.

Discussion

This experiment shows once again that subjects can learn to improve prediction accuracy in a novel complex causal environment simply by being exposed to input–output pairings, and this learning occurs for subjects exposed to input–output environments as well as to environments with an intervening variable. More important, the results for subjects in the intervening-variable group demonstrate that intervening concepts are spontaneously formed without being given explicit instruction that such a structure was possible or present. This result also extends our previous results by demonstrating (a) the formation of these concepts in an entirely new stimulus domain (cell systems rather than manufacturing systems), (b) a new training procedure (e.g., the use of causal systems with entirely different sets of weights), (c) the effects appear on the very first trial of each system (before any feedback about the system weights is given), (d) subjects do not carry over specific input–output associations from one system to the next, and finally, (e) subjects do not learn to focus on only one input variable.

Although we have shown that subjects will form these concepts spontaneously, they are not formed invariably. Instead, the tendency to form intervening concepts depends critically on the nature of the environment, the format of the inputs, and the amount of experience with that causal structure. Consistent with the third hypothesis outlined in the introduction, it seems that subjects begin without a dominant single intervening factor mediating their output predictions. For the input–output group, this state of affairs persisted across all training systems. For the intervening variable group, however, increasing exposure to the causal systems produced an increase in the amount of variance reproduced by the first principal component. Further, the format in which the inputs were presented strongly influenced the salience of the presence or absence of an intervening factor in the system. Specifically, inputs presented in an integral fashion encouraged the abstraction of an intervening concept more than inputs presented in a sequential, separate manner.

In sum, the major implication of these results is that subjects presented with a multiple input–output system in which an intervening factor is embedded learn something different than do subjects presented with a system without

an intervening factor. We have argued that what subjects learn is an intervening concept. If so, then they should be able to make inferences and extrapolate this knowledge to new situations. In the next two experiments, we test this implication.

These new experiments also bear on the issue of what classes of models might be best suited to account for the learning patterns evidenced. One might argue that the principal component results can be explained by either neural-network models (in which a hidden unit could function as an intervening concept) or exemplar-based models. In an exemplar model, subjects would learn to associate a particular set of outputs with a particular set of inputs. Instantiated within an adaptive-learning network in which there are generalization gradients of activation around each exemplar (i.e., set of inputs) (e.g., ALEX; Nosofsky & Kruschke, 1992), learned sets of exemplars (inputs) could then produce reasonable predictions for input sets that were slightly different from the learned sets. As training progressed through the first system and more input sets were learned, predictions would presumably become more accurate. For the intervening-concept system, this would also necessarily increase the proportion of variance explained by the first principal component (see Experiment 1 introduction). The key here is that if at the outset of the next system the learner carries over the input–output associations (exemplars) from the previous system, then a learner's predictions, though not accurate, would still be explained by the first principal component. In previous applications of exemplar models, previous associations have not been carried over from one learning task to an independent learning task in which completely new associations have to be learned (however, see Flannagan, Fried, & Holyoak, 1986, for evidence bearing on carryover effects). In this situation, it is not unreasonable to assume that subjects perceive some continuity across learning tasks (systems), and accordingly, subjects may initially carry over the learned associations to a subsequent system. Earlier, we reported an analysis that provided some evidence inconsistent with this carry-over hypothesis (see Experiment 2 results). However, this formulation (which essentially suggests that an intervening concept is not necessarily formed) is important and worthy of further examination in more incisive experiments reported next.

EXPERIMENT 3

In this experiment we used a transfer paradigm as an analytic technique to provide converging information on what is learned when subjects are exposed to a system in which outputs are mediated by an intervening factor. We first gave subjects extensive experience trying to predict the values for three outputs given a set of three inputs. As in previous experiments, subjects received one of two systems: an intervening-factor system or an input–output system. After receiving training on a system, subjects were given a transfer task in which a new fourth output was added to the system.

Consider that subjects have had extensive experience with a multiple input–

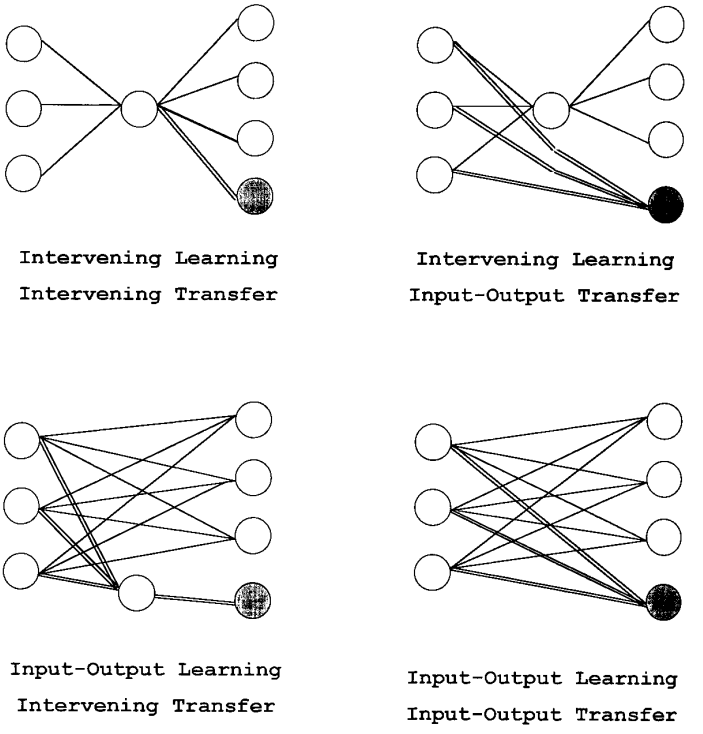


FIG. 4. A schematic diagram of four alternative transfer conditions in Experiment 3.

output system with an intervening factor and are now confronted with a new output for the system (not yet seen by the subjects). One possibility is that during training the subjects have learned that a hidden factor intervenes between a set of inputs and a set of outputs. If so, then two clear outcomes would be expected. First, if the new output is mediated by the same intervening factor as the old outputs, then these subjects should be more accurate at predicting the new output (because presumably the weights from the inputs to the intervening factor have already been at least partially learned) relative to subjects trained on an input–output system who are given a new fourth output with the same weight structure to predict. Second and importantly, this advantage for the intervening-variable subjects in predicting the new (fourth) output should not extend to a transfer situation where the new output is not mediated by the intervening factor (instead the output is mediated by each input separately, as in the input–output systems—Fig. 4 provides a schematic of these alternative transfer situations). That is, intervening-variable subjects transferred to the latter situation should perform significantly *worse* than similarly trained subjects transferred to the intervening variable-mediated output and they should perform no better than input–output subjects.

The alternative possibility is that subjects given the intervening system do not learn the intervening concept in the course of learning to predict the outputs. If so, then regardless of how the new output is related to the inputs, subjects trained on the intervening-factor system should show little advantage in learning to predict the new output relative to subjects trained on an input–output system. Note that this expectation holds even within the exemplar-based account sketched earlier. In an exemplar model, without direct experiences or associations with the fourth output, the model has no basis for making a reasonable prediction, with or without training on the intervening concept system.

Although our primary objective was to test the predictions just mentioned, we had a secondary interest in showing that the changes in cover story across Experiments 1 and 2 (factory and cell, respectively) did not play a prominent role in the emergence in Experiment 2 of a benefit in prediction accuracy (in terms of lowering error) for the intervening-factor subjects relative to the input–output subjects. Work in standard concept-learning paradigms has shown that the semantic context in which the concept task is presented can have powerful effects on the relative difficulty of learning different types of concepts (e.g., conjunctively vs disjunctively based concepts; Pazzani, 1991). This effect is attributed to the degree to which prior knowledge is activated by the learning context. Based on that research, it seems reasonable to wonder if the cell cover story with its concrete input dimensions induced more involvement of prior knowledge than the factory cover story, thereby facilitating the extraction and utilization of an intervening concept. Accordingly, in this experiment we used the first three training systems from Experiment 2 (three inputs and three outputs related by weight structures identical to those used in Experiment 2) and presented the systems in the same sequence. The only change was that the factory cover story was used instead of the cell cover story.

Method

Subjects and design. Twenty-seven undergraduate and graduate students attending Purdue University served as subjects, and as in the other experiments they were paid for their participation, None of the subjects had participated in the previous experiments. Subjects were randomly assigned to one of four experimental groups defined by the factorial combination of two between-subjects variables. One variable was the causal system structure used in training (intervening factor or input output), and the other was the relation of a new output (added during transfer) to the existing system structure. In one condition, the *intervening variable transfer* condition, the output was derived from the intervening factor of the intervening-variable system used in training. Thus, to obtain the values for the new output, a new output weight connecting the intervening variable to the new output was arbitrarily chosen and the output was directly computed from the preexisting input weight structure of

a trained intervening-variable system. Note that for the input–output training group, this transfer condition would not actually produce an intervening factor; it simply represents an input–output condition in which the values of the new output are exactly those assigned to the intervening variable transfer condition.

In the *input–output transfer* condition the new output was derived as in the input–output systems. Weights from the three inputs to the output were arbitrarily established (one weight per input–output link), and added to the system (either the input–output system or the intervening-factor system). Note that for the intervening-factor system, this transfer condition actually produced a hybrid system in which the three old outputs were derived from the intervening factor and the new output was derived directly from the new input–output weights added to the system (see Fig. 4 for a schematic).

Six subjects were assigned to the intervening-factor group given noncausal transfer and seven subjects were assigned to each of the three remaining groups.

Procedure. The instructions were similar to those used in Experiment 1, with the exception that the intervening-factor groups were neither given examples of systems with intervening constructs nor encouraged to try to form an intervening concept for the experimental tasks. Also, unlike in Experiment 1, subjects were given a hint that some weights would be negative (''Some inputs are good for output products and some other inputs are bad for output products''). After 5 practice trials, subjects received 50 trials on each of two systems (Systems 1 and 2 from Experiment 2), with a 15-min rest break between the systems. Subjects then returned the next day and received 100 trials on a third system (System 3 from Experiment 2), with a 15-min break after 50 trials. Finally, subjects returned a third consecutive day, and after receiving 50 more trials on the third system (and a 15-min break), they were presented with 50 transfer trials. During the transfer trials, subjects predicted four outputs (instead of three) from a set of three inputs and they received feedback. For these trials subjects were informed that the factory system worked the same as before except that it produced one more output product.

Stimuli. The first three training systems from Experiment 2 were used, but the factory cover story from Experiment 1 was substituted for the cell cover story. For transfer, the third training system was modified by adding a fourth output. In the intervening-variable transfer condition, the weight linking the intervening factor to the new output was .73. As mentioned above, the output value was computed by multiplying this weight by the value of the intervening factor computed from the system assigned to the intervening-variable training condition, and this value was then used for both input–output and the intervening-variable training conditions. In the input–output transfer condition, the three weights associated with the three new input–output links were .22, .70, and −.81. To obtain the output value for a given transfer trial, each weight was multiplied by the corresponding input value and the products summed.

TABLE 7

Mean Performance Measures (Mean Absolute Prediction Error, Correlation between Predicted and Actual Outputs, and Proportion Variance Reproduced by the First Principal Component for the First 25 Trials) for Experiment 3

Accuracy measures	Trial block	Intervening-variable	Input–output
MAE	1	15.64	20.01
	2	15.03	16.61
Correlation	1	0.25	0.26
	2	0.34	0.30
PC	System		
	1	0.58	0.57
	2	0.71	0.55
	3	0.74	0.58

Note. Each trial block represents 25 trials. MAE, mean absolute error; PC, proportion accounted for by the first principal component.

Results

The effect of training on prediction accuracy is reported first, followed by the more crucial analysis of performance on the transfer task. For all analyses the rejection level was set at .05.

Training. Because the first three sets of 50 trials were a conceptual replication of the previous experiment (they correspond to Systems 1–3, respectively, from Experiment 2), we first analyzed those training trials separately to examine whether patterns similar to those obtained in Experiment 2 emerged. Table 7 provides a summary of these results.

First, note that predictive accuracy based on the mean absolute error index improved across training blocks (25 trials per block) for both training groups, and according to both the mean absolute error index and the correlation index. This improvement was significant in the input–output training group, $F(1,13) = 121.97$, $MSe = 0.662$, $p < .0001$, but not for correlation ($F(1,13) = 1.58$, $MSe = 0.007$, $p < .2310$). In the intervening-training group, only for the correlation index did the improvement approach significance, $F(1,12) = 4.40$, $MSe = 0.012$, $p = .0577$.

More important, the first principal component accounted for most of the variance in the intervening variable subjects' predictions, and this value was significantly greater than that obtained for the input–output subjects, $F(1,25) = 7.31$, $MSe = 0.035$, $p = .0122$. This effect significantly interacted with system $F(2,50) = 6.05$, $MSe = 0.009$, $p = .0044$, reflecting that the difference in the groups in the amount of variance reproduced by the first principal component emerged only after the first training system. Planned comparisons confirmed that there was no group difference on the first system ($F < 1$),

TABLE 8
Mean Prediction Accuracy for the Final Training System in Experiment 3

Accuracy measure	Trial block	Intervening-variable	Input–output
MAE	1	14.43	16.04
	2	10.53	11.40
	3	11.42	11.82
	4	10.82	9.64
	5	10.85	10.89
	6	10.27	9.91
Correlation	1	0.21	0.39
	2	0.41	0.47
	3	0.34	0.43
	4	0.40	0.61
	5	0.34	0.53
	6	0.36	0.55

Note. Each trial block represents 25 trials. MAE, mean absolute error.

but there was a significant difference by the last system ($F(1,50) = 21.41$, $MSe = 0.009$).

Recall that System 3 was used for the transfer test, and subjects received a total of 150 trials of training with this system. Given the importance of performance on System 3, we also evaluated the effect of training block for this system separately for each training group. Table 8 shows that subjects in both groups improved their predictive accuracy with increased experience (i.e., across the six trial blocks) for both measures of accuracy. One-way ANOVAs with trial block as a within-subjects factor showed significant improvement in mean absolute error ($F(5,60) = 8.77$, $MSe = 3.51$, $p < .0001$, and $F(5,65) = 25.42$, $MSe = 2.97$, $p < .0001$ for intervening and input–output groups) and in the correlation measure for the input–output group, $F(5,65) = 10.12$, $MSe = 0.01$, $p < .0001$ ($F(5,60) = 2.28$, $MSe = 0.03$, $p < .0577$ for the intervening-variable group).

Transfer. Examination of individual performance during training revealed that individuals differed greatly in their ability to do the task. Accordingly, in line with other research on transfer (e.g., McDaniel & Schlager, 1990), we factored out individual ability differences for the fourth output by using each subject's mean absolute prediction error for the first three outputs during transfer as a covariate in the following transfer tests of the theoretically relevant contrasts outlined in the Introduction.

Two separate analyses of covariance (ANCOVA) tests were performed on the mean absolute error of the fourth output during the transfer test (see Table 9 for means). One test was conducted on the data from the intervening variable *transfer* group, and the other test was conducted on the data from the input–output *transfer* group. Note that for a given transfer group, exactly the same

TABLE 9
Mean Prediction Accuracy (Mean Absolute Prediction Error) for the Fourth (Transfer) Output (Experiment 3)

	Training condition			
	Intervening-variable		Input–output	
Trial block	Intervening transfer	Input–output transfer	Intervening transfer	Input–output transfer
1	15.32	25.78	19.54	25.55
2	16.34	25.07	21.02	22.57

Note. Each trial block represents 25 trials.

input–output pairs were presented to each training group during the transfer test on the new fourth output, so that direct comparisons across training groups are appropriate. The only difference between the two training groups for this test is what they learned during training.

The ANCOVA for the intervening-variable transfer groups revealed that the intervening-variable training subjects were more accurate in predicting the fourth output than the input–output training subjects, $F(1,10) = 5.11$, $MSe = 12.98$, $p = .0478$. This is consistent with the hypothesis that the subjects in the intervening-variable training group learned an intervening concept and used this concept to improve performance on the transfer test.

In contrast, the ANCOVA for the input–output transfer groups showed no differences between intervening-factor and input–output training subjects ($F < 1$).

Discussion

The results of the present experiment replicate the main findings of the previous experiment rather well, despite the change in cover story from biological cell to manufacturing systems. As in the previous experiment, a majority of the variance in the subjects' output predictions were reproduced by the first principal component for subjects trained with causal systems containing an intervening variable, but not for subjects trained with an input–output causal structure. Furthermore, this difference between the two groups increased as a function of increasing exposure to the causal structure. Finally, as in the second experiment, these results were obtained without any prior instructions regarding the presence or absence of an intervening factor. One interesting difference between the two experiments was that the magnitude of the first principal component for the first causal system was much larger under the biological cell cover story as compared to the manufacturing cover story (this was particularly true for the integral input format). This presumably

reflects both the effect of using a separable input format in the third experiment and the effect of prior knowledge due to changes in the cover story.

The transfer results provide the critical test of subjects' ability to apply a newly formed intervening concept to make new inferences. In this case, subjects were required to extrapolate their training with three earlier outputs to a new output variable. When subjects were required to predict a new fourth output that was related to the intervening variable (for the intervening-variable system), the subjects trained with an intervening-variable system showed a clear advantage over the subjects trained with an input–output system (in terms of being able to produce more accurate predictions). Because the advantage of the intervening-variable subjects relative to the input–output subjects was not found when the fourth output was not related to the intervening variable, the transfer advantage just described could not be explained by some sort of general facilitatory effect of having received intervening-variable systems (e.g., less fatigue, more motivation). Nor could the advantage be due to the possibility that more able learners were assigned to the intervening-variable group because the differences in transfer emerged in an analysis of covariance that statistically factored out learning ability (as assessed by final performance on the trained outputs). Instead, the transfer results imply that subjects who receive the intervening-variable systems acquired a different type of knowledge than did the subjects who received the input–output systems; specifically, the subjects receiving an intervening-variable system appear to have learned some sort of intervening concept or factor, and they were able to use this knowledge to extrapolate to a new output variable.

One last alternative possibility is that subjects in the intervening-variable training condition learned direct relations among the outputs themselves, without abstracting a mediating variable between inputs and outputs. Formulation of these direct relations would presumably be based on the salient perceptual information that is present when all (or nearly all) outputs are displayed simultaneously (as was the case in Experiments 1–3). That is, when all outputs are displayed simultaneously, the consistent graphic relations among the outputs (which would appear only for the intervening-variable training condition) might simply encourage knowledge of the direct relations displayed. Thus, it still is not entirely certain that subjects in the intervening-variable training condition have abstracted an intervening concept, and, accordingly, Experiment 4 was conducted to eliminate this alternative possibility.

EXPERIMENT 4

Our experimental strategy was to construct the learning and transfer context to eliminate the perceptual salience of the relations among outputs (in the intervening-training condition) and so that there would be no explicit or implicit suggestion to subjects that they even try to learn output relations. To this end, we presented subjects with only one output at a time during the

training phase so that no output ever appeared together with another output throughout the entire experiment. In this situation, we assumed there would be no motivation or encouragement for subjects to learn direct output relations. On the other hand, this paradigm would not necessarily discourage the formulation of an intervening concept.[5] Thus, if the intervening-variable training condition displays positive transfer to new input−output relations (relative to the input−output training condition), then based on the logic developed in Experiment 3, the strong implication is that subjects have abstracted an intervening concept.

Another purpose of this experiment was to extend the finding in Experiment 3 that intervening-concept training facilitated transfer to a new output mediated by the intervening concept. Acquisition of an intervening concept should also facilitate transfer when new inputs are added to the system. To demonstrate that positive transfer also can be obtained when a new input is added (rather than an output), subjects in this experiment received three inputs and four outputs during training, and then they were given an additional fourth input during transfer (see Fig. 5).

The design was also simplified from that in Experiment 3 in that incompatible training-transfer conditions were not included (e.g., intervening learning and input−output transfer). These conditions were used in Experiment 3 to establish that any positive transfer obtained in the intervening-variable learning condition was not due to possible general facilitative effects (e.g., motivation) associated with a particular type of training system. Having established that point, these conditions were no longer needed.

Method

Subjects and design. Nineteen undergraduate and graduate engineering students attending Purdue University participated. They were paid for their participation as in the other experiments. None of the subjects had participated in the previous experiments. Nine subjects were randomly assigned to the intervening-variable training condition and 10 subjects were randomly assigned to the input−output training condition.

Stimuli. Each subject received one system. The system contained three inputs and four outputs during the training phase, and an additional input was added to the system for the test phase. The input values were restricted and had three equally spaced values (0, 1, 2) during training and transfer. The weights were sampled from the interval [−1,+1]. The weights used for both training systems are given in Table 9. For the transfer phase, the weight of

[5] Note that in this experiment, once subjects had abstracted an intervening concept they may or may not also become aware of the relations among outputs. We make no claims about the extent to which subjects might be able to express knowledge of relations among outputs, once learning has occurred. The critical point is that the experimental paradigm discourages learning based only on noticing direct relations among outputs.

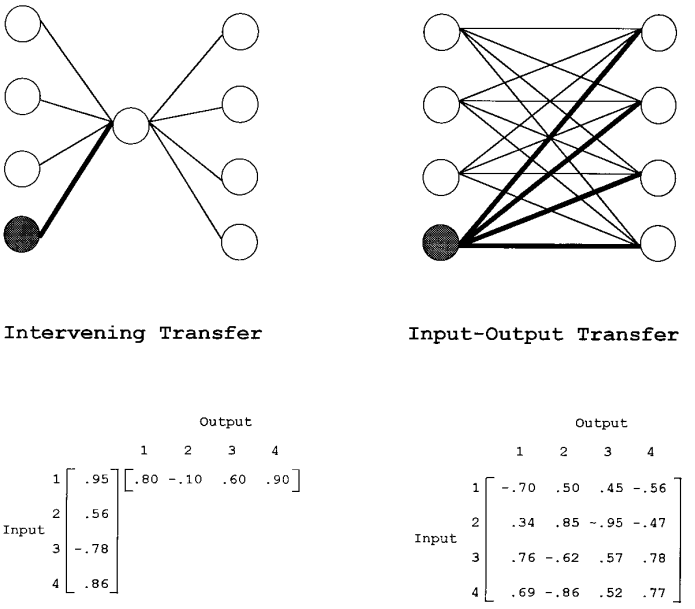


FIG. 5. A schematic diagram of transfer conditions in Experiment 4.

the newly added input to the intervening variable was arbitrarily chosen as .86 for the intervening-variable training condition. This new input weight multiplied by the weights of the outputs of the intervening-variable training condition were used as weights for the input−output condition (thus, the resulting weights connecting the new input to each of the outputs were 0.69, −0.86, 0.52, and 0.77). The remaining details for generating stimuli are the same as in Experiment 1.

Procedure. The factory cover story from Experiment 1 was used again. However, subjects were not given a hint that some weights would be negative and there were no practice trials, unlike Experiment 3.

During the training phase subjects received 312 trials, with two 1-min rest breaks after 108 and 216 trials, respectively. On each trial, subjects were given the three inputs simultaneously, but unlike the previous experiments, subjects made predictions about one output and then received the correct value for that output. The presentation order of the input sets was the same for the intervening-variable condition and the input−output condition, and we systematized the presentation order as follows: We constructed sequences of input sets such that the values for one input would vary across trials while the values for the other two inputs would remain fixed. For example, for the first trial the given set of inputs was (1, 0, 0) respectively, for the first, second, and third inputs, and for the second trial it was (2, 0, 0). A sequence was

TABLE 10
Presentation of Order of Inputs (Experiment 4)

	Training phase			
Trial	Input 1	Input 2	Input 3	Predicted output
1	1	0	0	Output 1
2	2	0	0	Output 1
3	1	0	0	Output 2
4	2	0	0	Output 2
5	1	0	0	Output 3
6	2	0	0	Output 3
7	1	0	0	Output 4
8	2	0	0	Output 4
9	0	1	0	Output 1
10	0	2	0	Output 1
11	0	1	0	Output 2
12	0	2	0	Output 2
13	0	1	0	Output 3
14	0	2	0	Output 3
15	0	1	0	Output 4
16	0	2	0	Output 4
17	0	0	1	Output 1
18	0	0	2	Output 1
19	0	0	1	Output 2
20	0	0	2	Output 2

formed by first exhausting the levels of the first input while subjects were required to make predictions about each output in turn. In other words, the two input trials just described were repeated four times in order for subjects to make predictions for each of the four outputs in turn. This sequencing produced the first eight trials. Then the values for the second input were varied, while those for the other inputs were fixed (i.e., this time the set of inputs had the values (0, 1, 0) and (0, 2, 0)). Again, these two input sets were repeated four times (once for each output), and so on for input 3. Table 10 illustrates part of the presentation order.

After the training phase, subjects received 48 transfer trials. At the beginning of the transfer phase, subjects were told that the factory system worked the same as before except that it contained one more input. The first 24 transfer trials required subjects to predict the value of the *first* output, given a set of four inputs (three existing and one newly added inputs). Feedback was given on these first 24 trials. The last 24 trials required subjects to predict the value of the other three outputs when all four inputs were given. For these trials, a set of eight input values was selected and subjects had to predict the second, third, and fourth outputs in turn (i.e., as in training, subjects made a prediction for only one output for each trial). No feedback was provided on these last 24 trials.

TABLE 11
Mean Prediction Accuracy (Mean Absolute Prediction Error and Correlation between Predicted and Actual Outputs) for Experiment 4

		Training condition	
Accuracy measure	Trial block	Intervening-variable	Input–output
MAE	1	23.08	22.18
	2	19.53	18.92
	3	19.06	17.95
Correlation	1	0.52	0.51
	2	0.58	0.72
	3	0.69	0.73

Note. The first and second blocks represent 108 trials, and the third block represents 96 trials.

Results

The effect of training on prediction accuracy is reported first, and then the more important analysis of performance on the transfer task follows. The major dependent measures were mean absolute error (MAE) and the correlation between subjects' predictions and the correct output. Analyses on principal component could not be performed, because only predictions about one output in one trial were made throughout the experiment, unlike previous experiments, which had multiple outputs. For all analyses the significance level was fixed at .05.

Training. As in the previous experiments, separate one-way repeated-measure ANOVAs (with training block as the independent factor) were conducted for each training condition and for each measure. Table 11 provides a summary of the means.

Predictive accuracy improved across the training blocks (108 trials for the first and second blocks and 96 trials for the third block) for both training groups according to both the mean absolute error index and the correlation index. In the intervening-variable group, the improvement was significant for the mean absolute error ($F(2,16) = 6.06$, $MSe = 7.18$, $p = .0110$) and for the correlation ($F(2,16) = 5.28$, $MSe = 0.014$, $p = .0174$). The improvement was also significant for the input–output group for the mean absolute error index ($F(2,18) = 4.46$, $MSe = 11.00$, $p = .0267$) and for the correlation index ($F(2,18) = 24.58$, $MSe = 0.006$, $p < .0001$).

Transfer. Recall that subjects received feedback on Output 1, and therefore no differences were expected on this output during transfer. The critical test is based on Outputs 2, 3, and 4. The difference in mean absolute error (averaged across Outputs 2, 3, and 4) was 39.5 (input–output) − 28.2 (intervening) = 11.32 favoring the intervening group. The difference in correlation (averaged across Outputs 2, 3, and 4) was .45 (intervening) − .07 (input–output)

TABLE 12
Mean Absolute Error and Correlations for New Fourth Input during Transfer in Experiment 4

	Output 1		Output 2		Output 3		Output 4	
Group	MAE	CORR	MAE	CORR	MAE	CORR	MAE	CORR
Intervening	10.9	.70	54.3	.15	14.8	.59	15.6	.62
Input–output	8.8	.79	46.4	.07	38.1	−.06	34.1	.31

= .38 favoring the intervening group. Table 12 shows the mean absolute errors and correlations from transfer for each output and each training group separately.

As in Experiment 3, we factored out individual differences for the transfer tests. Subjects' accuracy index on the first output during the transfer phase was used as a covariate. A multivariate analysis of variance test (MANOVA) was performed on the mean absolute error and the correlation measures for the second, third, and fourth output (i.e., the last 24 trials). The intervening-variable training subjects were significantly more accurate in predicting the second, the third, and the fourth output than the input–output training subjects for MAE ($F(3,13) = 6.64$, $p = .0059$), but not for the correlation ($F(3,13) < 1$).

The analysis based on MAE of each output showed that the intervening-variable training subjects did better than the input–output training subjects on the third ($F(1,15) = 29.27$, $MSe = 87.64$, $p < .0001$) and the fourth output ($F(1,15) = 7.71$, $MSe = 209.70$, $p = .0141$), but not on the second output ($F(1,15) < 1$). The correlation index on each output was significantly higher for the intervening-variable group compared to the input–output group on the third ($F(1,15) = 5.27$, $MSe = 0.38$, $p = .0365$), but not for the second output ($F(1,15) < 1$) or for the fourth output ($F(1,15) = 1.86$, MSe = 0.25, $p = .1926$).

In general, the intervening-variable group performed better than the input–output group during the critical-transfer test trials. The only exception to the general trend was for the mean absolute error for Output 2, and this difference was not significant. Moreover, the correlations for Output 2 show the opposite trend, which is superior performance for the intervening group. Recall that the relation between the new fourth input and Output 2 was negative, which caused very low performance for both groups. In sum, all of the significant differences were in the direction of better performance for the intervening-variable group.

Discussion

The results of the present experiment again provide evidence that subjects exposed to an intervening-variable training system learn something that facili-

tates transfer to new situations. One possibility mentioned at the conclusion of Experiment 3 was that subjects are learning output–output connections. In the present experiment, however, it is unlikely that intervening-variable training subjects were simply learning direct output–output relations. Only one output was given on each trial, thereby attenuating any demand that subjects compare outputs to one another and also eliminating the salience of the consistent perceptual pattern displayed when the outputs are considered together. Under these conditions it is reasonable to assume that the acquisition of output–output connections would be impaired or discouraged.

The alternative is that subjects are acquiring knowledge about an intervening factor that links the set of inputs to each output. This possibility is entirely compatible with the present training (and transfer) context. We would expect no interference with the acquisition of links between an intervening concept and each output if one output were presented at a time. Thus, the superior transfer performance of the intervening-variable training condition suggests that subjects abstracted an intervening concept, rather than acquired the direct output relations.

Another important point is that this pattern of results cannot be explained by the similarity of the final training weights and the transfer test weights. If similarity of the training weights and the test weights was the factor that affected subjects' performance, then the input–output group should have performed better than the intervening-variable group. As shown in Fig. 5, the similarity between the final training weights and the transfer test weights for the input–output condition is greater than that for the intervening-variable condition. This excludes the explanation that the intervening-variable group did better because they were exposed to a more familiar set of weights and were in an advantageous position. In fact, the opposite was true.

The transfer results provide the critical test of subjects' ability to apply a newly formed intervening concept to make new inferences. In this case, subjects were required to extrapolate their training with three earlier inputs to a new input variable. When subjects were required to predict using a new fourth input that was related to the intervening variable (for the intervening-variable system), the subjects trained with an intervening-variable system showed a clear advantage over the subjects trained with an input–output system (in terms of being able to produce more accurate predictions). The transfer results imply that subjects who receive the intervening-variable systems acquired a different type of knowledge than did the subjects who received the input–output systems; specifically, the subjects receiving an intervening-variable system appeared to have learned some sort of intervening concept or factor, and they were able to use this knowledge to extrapolate to a new input variable.

GENERAL DISCUSSION

The preceding experiments provide converging evidence that individuals spontaneously detect and use intervening concepts when confronted with

novel causal environments that actually contain a hidden intervening factor. This is a fundamental finding about human conceptual learning because the presence of a causal environment does not obligate individuals to learn an intervening concept. In these experiments, subjects could instead have learned a set of individual input–output connections, or they could have learned only about the correlational relations among the outputs (relations that must emerge when the outputs are mediated by an intervening factor). The results, especially of Experiments 3 and 4, suggest that subjects' experiences with causal environments lead to something other than learning just input–output associations or just the direct relations among the outputs.

Having established that subjects did use intervening factors, we are now faced with two interdependent questions: What is it that subjects actually learn about the intervening factor, and how is this knowledge learned? We consider two theoretical alternatives for addressing these questions.

Exemplar-Based Approach

Consider one subject's verbal report on the postexperimental questionnaire from the intervening-variable group in Experiment 2: ''I tried [sic] to remember other inputs (size, shape, etc.) and then I compared them. I just tried [sic] to use recall to do this lab.'' This low level of understanding of the input–output relations appears to reflect an exemplar-based learning process wherein new output predictions were based on recall of previous associations between similar inputs and outputs. As discussed earlier, the consistent finding of a large principal component at the beginning of training on each new system does not necessarily discredit this theoretical position. In principle, a subject that learns according to an exemplar-based model would be able to learn the appropriate outputs associated with each input set by the end of training, and would thereby produce predicted outputs that reflect the output correlations produced by the intervening-variable system. The critical finding of a large principal component at the beginning of training (for systems beyond the first training system) could be explained by the addition of a relatively straightforward assumption. The assumption is that subjects generalize their associations learned at the end of training on one system to the beginning of training on the next system. Therefore, subjects' predicted outputs at the beginning of each new system, though not necessarily accurate, would be correlated with each other, and this would produce a large first principal component.

This explanation was examined in Experiment 2 by correlating the average of subjects' predictions for the first 10 stimuli of each new system with the outputs that would be produced from those same stimuli generated by the *previous system.* This produced an average correlation of $r = .06$, which is quite low compared to the $r = .86$ correlation between the subjects' outputs and the correct outputs from the last 10 stimuli of the same system.

This explanation encounters more serious difficulties with the transfer performance from Experiments 3 and 4. For example, consider transfer to the

new fourth output in Experiment 3. Even after extensive training on the first three inputs and three outputs, there would be no associations between the inputs and a new fourth output presented for the first time during transfer in Experiment 3. That is, with this type of exemplar-based representation, there would be no prior information on which to base predictions initially for the new fourth output. Consequently, in contrast to the obtained results, all groups would be expected to perform equally well on the fourth output. Similarly, for Experiment 4, according to the exemplar approach, there would be no associations between the new fourth input variable and the last three outputs during transfer from the new fourth input. Therefore, an exemplar-based level of understanding is not sufficient to capture what was learned in Experiment 4 (see the Appendix for a more rigorous argument).

Hidden-Unit Approach

Another idea is that subjects form a multilayer associative network (Rumelhart & McClelland, 1986) to solve this learning problem. In this representation, the hidden layer plays the role of the intervening factor by mediating between the inputs and the outputs of the network. There are two possible explanations for how this representation might be used to support learning in the current context.

The first explanation is that the intervening factor is essentially captured in the weights of the distributed representation involving the entire ensemble of hidden units. If so, for the intervening factor to emerge when a new system is initially encountered (as indicated by the principal component analyses), the entire weight structure from the previously learned system must be carried over and applied to the new system.

This explanation is virtually the same as that previously described for the exemplar approach. However, as we pointed out earlier, this explanation is inconsistent with the fact that subjects' predictions at the beginning of each new system were uncorrelated with the corresponding outputs produced by the previous system. Also, for the same reasons that we mentioned for the exemplar approach, this explanation cannot account for the transfer results from Experiments 3 and 4. These experiments indicate that the intervening factor that emerged in the present experiments was not mediated by a distributed weight structure learned on a previous system and applied initially to the next system.

Another possible explanation for how an associative-network model could account for the learning observed herein is that, as training proceeds, subjects in the intervening-factor group somehow (specified below) form a single hidden unit that intervenes between the input and output nodes of the network. This representation captures the essence of the intervening concept, and thereby reflects a deeper level of understanding than does the exemplar-based approach.

On close inspection, the results of our study pose a number of new chal-

lenges for multilayer network learning models. First, an important issue that still needs to be addressed is how subjects come to form a single hidden-unit architecture. There are at least two alternatives. One is that subjects begin with a single hidden-unit architecture and then add more hidden units (essentially, they change weights from zero to nonzero values) as needed to improve accuracy (e.g., in the case of the input–output systems, subjects would need to add hidden units to attain optimal performance). A second possibility is that subjects begin with a large number of hidden units and then eliminate as many as possible to simplify the network. For the intervening-variable group the simplest architecture would be a single hidden-unit network. Contrary to the first possibility and consistent with the second, the results of Experiments 1, 2, and 3 showed that subjects did not begin the experiment using a single intervening concept to generate their predictions. The first principal component started out low for the first few systems and then gradually increased to a high value for the intervening-variable group.

Given that it appears that learning of the intervening concept is achieved by simplification, the question that follows is what type of learning algorithm could produce this simplification. The standard back propagation learning algorithm embedded in multilayer networks (Rumelhart, Hinton, & Williams, 1986) fails to provide any means for generating parsimonious architectures. We therefore have developed an extension to the standard algorithm that is sensitive both to accuracy and to parsimony, and we use this algorithm as the foundation for a model of intervening-concept learning. The model, as well as simulations demonstrating its utility, is described next.

Hidden-Unit Parsimony Model

The main feature of our model is a gradient-descent learning algorithm in which the objective function is a weighted average of two indices, an accuracy index, and a parsimony index.[6] The accuracy index is based on the mean absolute prediction error. The parsimony index is based on the magnitude of the weights associated with the hidden units. More specifically, nonzero weights associated with all but one particular hidden unit are penalized. Space limitations do not permit a full exposition of the model here, but a complete description of the model appears in Busemeyer, McDanial, and Byun (1996). The appendix describes a short and simple version of the model for the illustrative purposes of this article.

We conducted simulations of the model applied to the multiple input–output environments used in Experiments 2 and 3. Accordingly, there were

[6] The idea of using simplification as part of the learning objective is not so new, and previous theorists have proposed similar ideas in other contexts. (See, e.g., Kruschke & Movellen, 1991, and references therein. Also note that Rumelhart, 1988, proposed an index of parsimony similar to the one we used.) We are not committed to one specific approach, but instead we use a specific approach to show that the general principle does work.

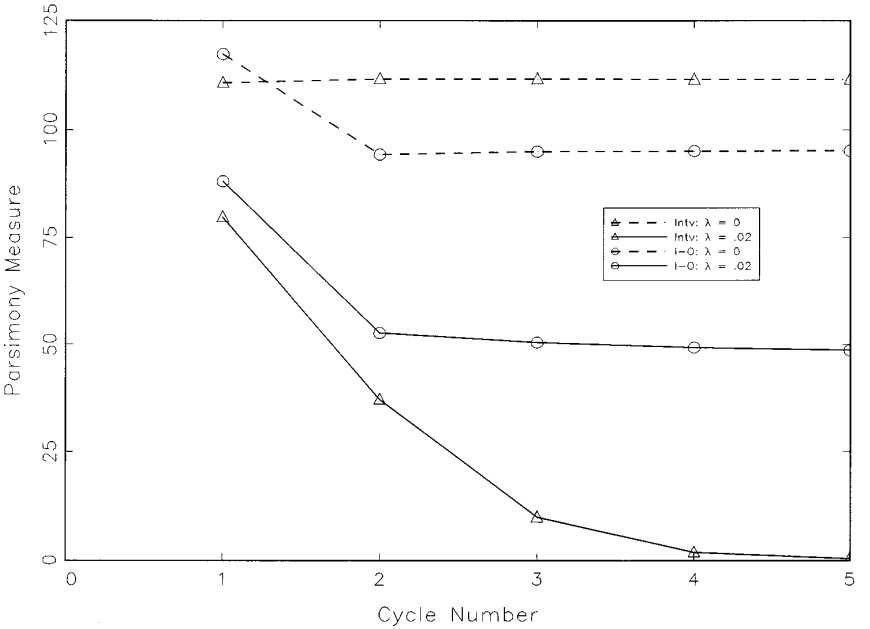


FIG. 6. Values of the parsimony index for the model simulations.

three input nodes, three hidden-unit nodes, and three output nodes. Two parameters were involved in the simulation—a learning-rate parameter and a parameter that weighted the contribution of the parsimony index to the objective function. The learning-rate parameter was arbitrarily set at .45, and the weight parameter was varied across simulations. In some simulations the weight parameter was 0, thereby excluding parsimony from the objective function, and in other simulations the weight parameter was .02. The training was patterned after that experienced by subjects in Experiment 3, such that the model received several cycles of 50 trials of training on a system (such as subjects received in system 3 prior to transfer). This allowed us not only to compare the model's performance against the subjects' patterns in training, but also to test how well the model (qualitatively) accounted for the transfer patterns. In considering training, the resultant network structure after training as a function of the involvement of parsimony during learning was of interest. The parsimony index served as an indicator of structure, with a value of 0 obtained when one and only one hidden unit had nonzero weights (with the input and output nodes). Values above 0 represented situations in which weights associated with other hidden units became increasingly nonzero.

Figure 6 shows the parsimony values of the model as a function of training on an input–output system with and without an intervening factor, and as a

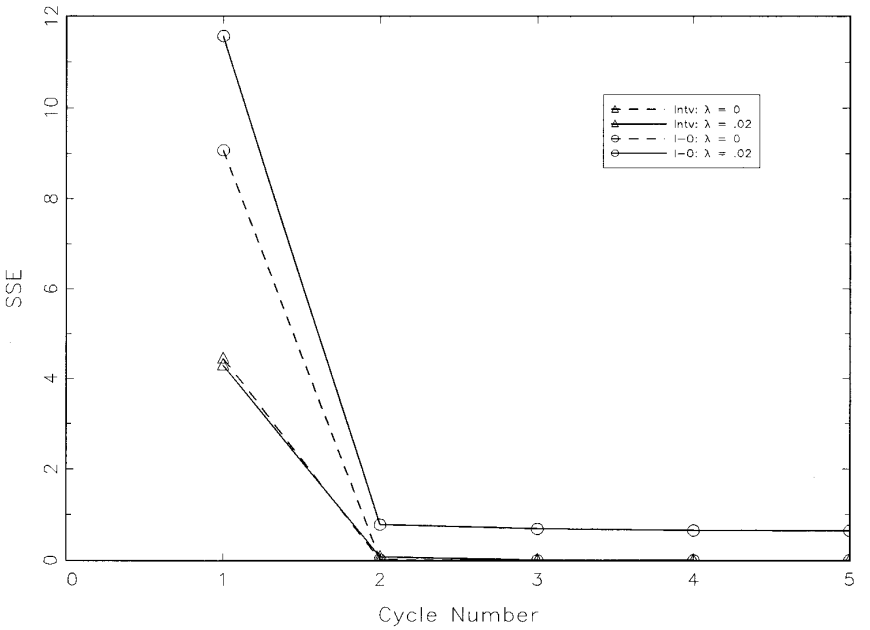


FIG. 7. Values of prediction error (SSE index) for the model simulations.

function of whether parsimony was included in the objective function (intact lines) or not (dashed lines). Examination of Fig. 6 shows that when the model is presented with an intervening-factor system and parsimony is included in the objective function, as learning proceeds a simplified network evolves such that finally only one hidden unit is functional (i.e., the parsimony index drops to 0). The other entries in Fig. 6 reveal that simplification is not an inherent property of the network model. When parsimony is not included in the objective function, the network asymptotes at a solution that is not reflective of a single intervening factor. A further test of the model is to examine its response when it is exposed to an input–output system without an intervening-concept structure. In this case, even when parsimony is included in the objective function, the model does not arrive at a solution reflective of a single intervening factor. Instead, as with the human learners in this study, the model was sensitive to the structure of the environmental system. In addition, it is important that when the model simplified the network in learning to predict the outputs for a system with an intervening factor, such simplification was not produced in lieu of accuracy. Figure 7 shows the mean prediction error for the simulations, and as can be seen, the prediction accuracy for the simulation with parsimony included improves substantially with training. This pattern of improvement compares favorably with that displayed by subjects in Experiment 3 (given repeated training on the same system; see Table 8).

The model also fares well in simulating the observed transfer to a new fourth output (in Experiment 3). At the conclusion of training, as described above (and summarized in Fig. 6 and 7), the model (the simulation with the parsimony-weighted objective function) was given 50 trials in which it was required to predict a new fourth output in addition to predicting the three outputs on which it had been trained. In the simulation, at the outset of transfer a fourth output node was added to the network, and the weight(s) to that node were set at 0. When the fourth output was constructed from the intervening factor used in the training system (for the intervening-factor condition), the mean prediction error for the fourth output was substantially lower (.54) after the simulation was trained on the intervening-factor system than after the simulation was trained on the input–output system with no intervening factor (4.49). This pattern mimics (qualitatively) the result obtained with the human learners. The model's advantage after training on an intervening-factor system was eliminated (just as for the human learners) when the model was transferred to a fourth output constructed from three new individual links to each input (the input–output transfer condition in Experiment 3). In this case the mean prediction error for the fourth output was somewhat higher (6.74) after training on an intervening factor system than after training on a system with no intervening factor (4.54). Learning to predict a new fourth output was facilitated for the model when it was trained and transferred to intervening-variable systems presumably because once the architecture had been simplified to a single hidden unit, the model only had to learn one output relation rather than three input–output relations (which occurs in the other training-transfer situations).

In conclusion, the empirical and theoretical work reported in this article has broken new ground on questions concerning the formation of intervening concepts and on extending adaptive network models to account for learning in these contexts. The experiments suggested that if the multiple input–output environment is structured with an intervening variable, then learners will detect this structure by forming an intervening concept. Moreover, they exploit the concept by extrapolating it to new input–output relationships. The implication here is that the ability to learn intervening concepts is a commonplace cognitive skill, not one restricted to a few exceptional individuals. Further, this cognitive skill appears to be captured by a class of learning model that has been applied to a wide range of learning tasks. Perhaps more importantly, the model briefly mentioned (see Busemeyer et al., 1996, for more details) is the first *working* model of which we are aware (published) that has successfully incorporated the pressures of parsimony into the learning process to form a single intervening factor. Our instantiation of parsimony is clearly closely tied to the present learning context; nevertheless, the simulations provide evidence that an adaptive network model is sufficient, with certain assumptions, to account for intervening-concept learning. Given a network model, a model for which learning would proceed at a relatively automatic

and/or nonconsciousness level, one implication is that the level of understanding for most subjects is such that they may not be able to coherently express the concept. It may be left to exceptional individuals to achieve a level of understanding that allows an explicit, articulate expression of the concept. Future research on intervening-concept learning may shed more light on how, for example, Newton discovered gravity, Mendel discovered the gene, and Spearman discovered general intelligence.

APPENDIX

This appendix has two parts: The first part identifies precisely how Nosofsky and Kruschke's (1992) ALEX model of categorization fails to account for the transfer results reported in Experiment 3, and the second part provides a mathematical description of the hidden-unit parsimony model used to learn a single intervening variable within a standard single hidden-layer architecture.

1. Transfer with ALEX

According to the ALEX model, each input pattern produces a distribution of activation in an n-dimensional perceptual space. Note that this input activation is generated by filtering the input stimulus, where the filters control the amount of attention allocated to each stimulus dimension. However, the attentional properties of the model are *not* relevant to the argument below, so this aspect is not described in detail.

The n-dimensional column vector $A(t)$ will be used to denote the distribution of activation produced by the input pattern presented on trial t (regardless of how attention is allocated). This activation is then mapped into an output-activation vector, symbolized by the m-dimensional column vector $R(t)$, where each coordinate of $R(t)$ represents one of the response categories. The mapping from the input activation to the output activation is accomplished by an $(m \times n)$ associative weight matrix as follows: $R(t) = W(t)A(t)$. The associative-weight matrix is updated following feedback on each trial by using an m-dimensional column vector, $Z(t)$, containing zeros everywhere except for the correct response categories, which receive values of 1.0. The weight matrix is modified according to the delta learning algorithm, $W(t) = W(t - 1) + \alpha \cdot [R(t) - Z(t)] \cdot A(t)^{\mathrm{T}}$, where the superscript T indicates the transpose operation. The weight matrix is initially set to 0. (This describes the associative-learning part of Kruschke's model, which is the part that is relevant to the present argument. The attentional-learning part is not relevant because the same argument applies for any allocation of attention.)

Note that the rows of the weight matrix W correspond to the output response categories, with one row for each unique response category. Suppose that the weight matrix W is arranged so that rows l to p correspond to all of the combinations of output-response categories possible

for the three output-response dimensions that received feedback during training for Experiment 3. The remaining m−p rows correspond to all of the output response categories possible for the new fourth output introduced for the first time during transfer for Experiment 3. According to the learning algorithm used by ALEX, all of the weights connecting the input-activation nodes to the last m−p output-response nodes remain at the initial weight values (0), because no feedback was presented during training that would change these weight values (i.e., the error signal, $R(t) - Z(t)$, is always 0 for the last m−p coordinates during training). This would be true for both the group trained with an input−output structure and the group trained with an intervening-concept structure. During transfer, both training groups receive exactly the same input stimuli and feedback. Therefore, ALEX predicts no differences in transfer performance for these two training groups. A similar argument can be made to show how the existing version of ALEX cannot explain the results for Experiment 4. ALEX was never designed for application to learning tasks that introduce new input dimensions or new output categories. Thus these results do not cast doubt on the validity of ALEX for its originally intended domain of application. Also, this is not to say that there is no way to extend the model to account for new input dimensions or output-response categories never seen before (see Delosh, Busemeyer, & McDaniel, in press, for example). We simply argue that the existing model cannot explain the current results.

2. Hidden-Unit Parsimony Model

Consider a standard single hidden layer connectionistic network architecture with n input nodes, m hidden nodes, and p output nodes. The weight $w_{1,ij}$ of the $m \times n$ weight matrix W_1 represents the weight connecting hidden-unit i to input-unit j. The weight $w_{2,jk}$ of the $p \times m$ weight matrix W_2 represents the weight connecting output node j to hidden-unit node k. The weight matrices are updated according to the following learning algorithm:

$$W_1(t) - W_1(t-1) = \alpha \cdot \delta_1(t) - \lambda \cdot \theta_1(t),$$

$$W_2(t) - W_2(t-1) = \alpha \cdot \delta_2(t) - \lambda \cdot \theta_2(t).$$

The first component on the right-hand side, $\delta_i(t)$, is the standard error correction following feedback on trial t that is typically used in the delta or generalized delta learning rule (cf. Rumelhart, Hinton, & Williams, 1986). It is well known that this first component is designed to minimize sum of squared prediction error.

The second component on the right-hand side, $\theta_i(t)$, is a type of parsimony index. It is designed to minimize deviations from a single intervening-concept representation of the associative relationships. More formally, we define a measure of parsimony separately for each weight matrix as

$$\gamma_1(t) = \sum_{i=2,m} \sum_{j=1,n} w_{1,ij}^2,$$

$$\gamma_2(t) = \sum_{j=1,n} \sum_{k=2,p} w_{2,jk}^2.$$

The first index penalizes for employing weights linking inputs to any hidden unit other than the first. The second index penalizes for employing weights linking any hidden unit other than the first to each output. The gradients of each of these two parsimony indices are then used to influence the direction of learning:

$$\theta_{1,ij}(t) = \partial\gamma_1(t)/\partial w_{1,ij},$$
$$= 2 \cdot w_{1,ij} \text{ if } i > 1, \qquad \text{otherwise } 0.$$
$$\theta_{2,jk}(t) = \partial\gamma_2(t)/\partial w_{2,jk},$$
$$= 2 \cdot w_{2,jk} \text{ if } k > 1, \qquad \text{otherwise } 0.$$

In the simulation reported under General Discussion, we employed $n = m = p = 3$ for simplicity. Also for simplicity, linear hidden-unit nodes were employed, because this was all that was necessary for the linear-prediction task that subjects were asked to learn in these experiments. The inputs to the network were simply equated with the input stimulus magnitudes used in Experiment 3.

In addition to the learning algorithm, we assume that at the outset learners adopt a simple architecture in which each input is connected to its own hidden unit, which in turn is connected to one output. In other words, the initial weight matrices were set equal to diagonal matrices. This assumption is consistent with responses on the postexperimental questionnaire in which some subjects stated that they initially tried to relate one input to one output (cf. first paragraph under General Discussion), and the assumption is also in line with research using single-output tasks indicating that subjects display an initial bias toward simple solutions (e.g., Brehmer, 1974; Koh & Meyer, 1991).

REFERENCES

Bentler, P. E. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, **31,** 419–456.

Bourne, L. E., Jr. (1966). *Human conceptual behavior*. Boston: Allyn and Bacon.

Brehmer, B. (1974). Single-cue probability learning as a function of the sign and magnitude of the correlation between cue and criterion. *Organizational Behavior and Human Performance*, **9,** 377–395.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New York: Wiley.

Busemeyer, J. R., McDanial, M. A., & Byun, E. (1996). The use of intervening variables in causal learning. In Shanks, D., Holyoak, K. S., & Medin, D. (Eds.), *Psychology of learning and motivation: Vol. 34. Causal learning.* San Diego: Academic Press.

Delosh, E. L., Bussemeyer, J. R., & McDaniel, M. A. (in press). Extrapolation: The sine qua

non for function learning. *Journal of Experimental Psychology: Learning, Memory , and Cognition.*

Flannagan, M. J., Fried, L. S., & Holyoak, K. S. (1986). Distributional expectations and the induction of category structure. *Journal of Experimental Psychology: Learning, Memory, Cognition,* **12,** 241–256.

Garner, W. R., Hake, H. W., & Erickson, C. W. (1956) Operationism and the concept of perception. *Psychological Review,* **63,** 149–159.

Hull, C. L. (1920). Quantitative aspects of the evolution of concepts. *Psychological Monographs,* **28** (Whole No. 123).

Hunt, E. B. (1962). *Concept learning: An information processing problem.* New York: Wiley.

Hunt, E. B., Martin, J., & Stone, P. (1966). *Experiments in induction.* New York: Academic Press.

Johnson-Laird, P. N. (1983). *Mental model: Toward a cognitive science of language, inference, and consciousness.* Cambridge, MA: Harvard 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., & Movellen J. R. (1991). Benefits of gain: Speeded learning and minimal hidden layers in back-propagation networks. *IEEE Transaction on Systems, Man and Cybernetics,* **21,** 273–280.

Levine, M. (1975). *A cognitive theory of learning.* Hillsdale, NJ: Lawrence Erlbaum.

McDaniel, M. A., & Schlager, M. S. (1990). Discovery learning and transfer of problem solving skills. *Cognitive and Instruction,* **7,** 129–159.

Miller, N. E. (1959). Liberalization of basic S-R concepts: Extensions to conflict behavior, motivation, and social learning. In S. Koch (Ed.), *A study of a science* (Vol. 2, pp. 196–292). New York: McGraw-Hill.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review,* **92,** 289–316.

Neisser, U. (1987). From direct perception to conceptual structure. In U. Neisser (Ed.), *Concepts reconsidered: The ecological and intellectual bases of categories.* Cambridge: Harvard University Press.

Nosofsky, R. M., & Kruschke, J. K. (1992). Investigations of an exemplar-based connectionist model of category learning. In D. Medin (Ed.), *The psychology of learning and motivation.* New York: Academic Press.

Qin, Y., & Simon, H. A. (1990). Laboratory replication of scientific discovery processes. *Cognitive Science,* **14,** 281–312.

Rosch, E., & Lloyd, B. (1978). *Cognition and categorization.* Hillsdale, NJ: Erlbaum.

Rumelhart, D. E. (1988). *Brain style computation.* Paper presented at the 21st annual meeting of the Society for Mathematical Psychology, Northwestern University, Evanston, IL.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations.* Cambridge: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations.* Cambridge, MA: MIT Press.

Smith, E. E., & Medin, D. L. (1981). *Categorization and concepts.* Cambridge: Harvard University Press.

Tatsuoka, M. (1988). *Multivariate analysis: Techniques for educational and psychological research.* New York: Macmillan.

(Accepted February 6, 1996)